UNIVERSIDAD CARLOS III DE MADRID

Escuela Politécnica Superior

Proyecto Fin de Carrera

Análisis de Herramientas de Detección de Idiomas en Facebook

Ponente: Estela Sánchez Delgado

Tutor: Ángel Cuevas

Leganés, Octubre 2015

PROYECTO FIN DE CARRERA:

Análisis de Herramientas de Detección de Idiomas en Facebook

Autor: Estela Sánchez Delgado

Tutor: Ángel Cuevas Rumin

Departamento: Ingeniería Telemática

Miembros del Tribunal:

Presidente: Manuel Urueña

Secretario: Raquel Aparicio

Vocal: Sara Tena

Fecha de lectura: 2 de Octubre 2015

Calificación:

"La constancia es la virtud por la que todas las cosas dan su fruto."

Arturo Graf

AGRADECIMIETOS

Me gustaría darle especialmente las gracias mi tutor, por haber dedicado tiempo y esfuerzo en este proyecto para terminarlo justo a tiempo.

También a mi familia y a Pablo por su apoyo incondicional cada día y por hacer que no tirara la toalla en ningún momento.

RESUMEN

En este proyecto fin de carrera se pretende realizar un análisis de las herramientas de detección de idiomas y testarlas para los comentarios de Facebook cuya particularidad es que son textos muy cortos y con lenguaje coloquial.

Se realizará un análisis de las herramientas del mercado y se automatizará aquella que de mejores resultados para el tipo de mensajes seleccionados.

En el resto de capítulos se evaluará el rendimiento para el caso práctico y se mostrará la utilidad de esta herramienta para las empresas.

INDICE

ÍNDICE GRÁFICAS

ÍNDICE FIGURAS

CAPÍTULO I: INTRODUCCIÓN Y MOTIVACIÓN

Este Capítulo contiene un primer apartado de Introducción en el que se da una visión general de lo que se va a tratar en el proyecto y en el segundo apartado se describe la necesidad y cuál es el objetivo del proyecto.

I.1 Introducción

Actualmente, las redes sociales son unas de las principales herramientas de comunicación utilizadas por los internautas y se han convertido en una parte vital de las relaciones entre amigos, familiares y desconocidos. Están sustituyendo a otras herramientas de comunicación como el correo electrónico o la mensajería instantánea.

En los últimos años se ha producido un incremento de usuarios debido a la evolución de las redes sociales, ya que pueden servir tanto para que los usuarios se comuniquen, compartan contenidos, ideas... como para que las empresas se promocionen o mejoren su imagen, y la comunicación con sus clientes sea más fluida.

En el estudio de Social Media 2015 realizado por Online Business School (OBS), indica que el 73% de los internautas de España (17M usuarios) utiliza de forma activa las redes sociales en 2014 y sólo el 8% no disponen de cuenta en ninguna red social.

Según el estudio las tres redes más usadas en 2014 por los internautas españoles son Facebook, Google+ y Twitter. El 88% de los españoles que utilizan Internet tiene cuenta en Facebook (frente al 87% en 2013), el 59% en Google+ (56% en 2013) y el 56% en Twitter (54% en 2013).

Este uso de las redes sociales no es sólo a nivel nacional sino que está extendido mundialmente, lo que supone una interconexión entre personas de cualquier parte del mundo y para las empresas poder llevar su marca de manera sencilla a cualquier lugar. Facebook es líder en volumen de usuarios en todas las regiones del mundo, seguida de Google + con excepción de Norte América, que es YouTube la segunda opción.

En el ámbito profesional, las redes sociales también tienen gran relevancia para la búsqueda o promoción de empleos, según OBS el 32% de usuarios españoles utilizan LinkedIn que es la principal red profesional. El otro principal uso para las empresas es como medio publicitario, en el mismo estudio el 23% de los usuarios de redes sociales es seguidor activo de sus marcas preferidas en 2014.

Independientemente del tamaño y antigüedad de las empresas las redes sociales se han convertido en un nexo entre las empresas y los clientes. Es la clave del éxito de un plan de marketing digital ya que de manera rápida se puede llegar a un gran volumen de personas y a bajo coste se puede promocionar una marca o producto incrementando así las ventas.

En este contexto se presenta el proyecto fin de carrera "Análisis de Herramientas de Detección de Idiomas en Facebook".

Debido a que los comentarios de las redes sociales tienen un lenguaje particular, los usuarios utilizan textos cortos, con abreviaturas, expresiones coloquiales... se pretende realizar un análisis detallado de las herramientas existentes en el mercado para evaluar el funcionamiento de las misma con el caso particular del lenguaje que se usa en las redes sociales, concretamente en Facebook. En base a la mejor solución que se encuentre en el mercado se proporcionará en este proyecto de fin de carrera una herramienta automática que clasificará el idioma de todos los comentarios. Esto permitirá a las empresas obtener información adicional de los comentarios en su página de Facebook y poder evaluar si sus campañas publicitarias tienen el efecto deseado o si están enfocadas en el idioma correcto.

En el Capítulo III se ofrecerá una visión general de las herramientas que existen en el mercado con sus ventajas y limitaciones.

En el resto de capítulos se profundizará en el análisis de las herramientas seleccionadas, acorde al caso de uso que concierne este proyecto, se mostrará las problemáticas de cada herramienta y se evaluará el rendimiento. Concretamente en el capítulo IV se evaluará el rendimiento de la short list y se elegirá la herramienta óptima para automatizar.

En el Capítulo V y VI, se mostrará la utilidad y la usabilidad de la herramienta y los costes de ponerla en producción.

Adicionalmente se adjunta los textos utilizados para los estudios, el código del proyecto y las referencias.

I.2 Motivación del Proyecto Fin Carrera

El objetivo es proporcionar a las empresas un mecanismo de mejora continua para sus campañas de Marketing Digital. Es clave en este contexto social donde las redes sociales han tomado mucha importancia y tienen una gran penetración en la población, recopilar todo el feedback de los usuarios y utilizarlo para mejorar en el futuro y así estar alineado con las necesidades de los clientes.

Las campañas de Marketing Digital tienen en cuenta muchos aspectos, uno de ellos es decidir en qué idioma o idiomas se lanzan y si alcanzan al público que se ha definido en la campaña. Con este proyecto se pretende dar una herramienta automática a las empresas para que puedan evaluar las respuestas de los usuarios a un anuncio, comunicado, promoción... publicado en la red social Facebook, concretamente permite evaluar si el idioma en el que se ha lanzado la campaña se corresponde con el idioma de los post de los usuarios o si en determinadas empresas o páginas de Facebook un idioma causa una mayor respuesta que en otras, permitiendo así estudiar el comportamiento de los usuarios y poder centrar el idioma de las campañas según el público objetivo que se quiera alcanzar.

En el mercado hay múltiples soluciones que permiten obtener el idioma de un texto, estas soluciones se basan en algoritmos especiales y estadísticos para analizar las combinaciones de letras en cada oración, y determinar así el idioma de palabras individuales o frases. Muchos de estos algoritmos usan un modelo vectorial para determinar el idioma y algunas aplicaciones públicas como los servicios que brinda Google hacen uso evidente de este tipo de clasificación.

Se ha realizado un análisis exhaustivo de dichas soluciones eligiendo la que mejor se adapta a este tipo de textos, ya que en las redes sociales como se ha comentado en el apartado anterior se utiliza un lenguaje coloquial y los post suelen ser cortos y con abreviaturas. El valor añadido que se ha aportado a la solución del mercado ha sido la automatización de los textos, es decir poder gestionar todos los post al mismo tiempo sin tener que analizarlos manualmente uno a uno. Por tanto, se ha creado una herramienta para que las empresas puedan obtener el idioma de todos los post a la vez y así poder realizar análisis posteriores sencillos.

Finalmente, se introduce un caso práctico en el que se muestra un ejemplo de uso de la herramienta, la información que se puede obtener y el uso que se le puede dar. Se quiere mostrar el potencial de la herramienta y la utilidad de la misma.

CAPÍTULO II: CONTEXTO

Este Capítulo antes de adentrarnos en el análisis de las herramientas se muestra en qué contexto surge el proyecto ya que es necesario saber cómo han evolucionado las redes sociales, el cambio que han supuesto en la sociedad y su importancia, debido a que las personas dedican una gran parte del día a visitar, subir contenidos o buscar información en ellas.

II.1 ¿Qué entendemos por redes sociales

De una manera sencilla, una red social es una estructura en forma de sitio web en la cual se agrupan los usuarios según una serie de criterios que ellos mismos eligen. Estos criterios pueden variar mucho, desde actividades específicas como la gastronomía o los vehículos personalizados, o sitios como Facebook, en el cual se agrupan personas conectadas por su parentesco o amistad.

El objetivo de estas redes sociales es que sus usuarios establezcan lazos entre ellos, facilitando la comunicación y el intercambio de conocimientos e información de manera virtual, sin importar el lugar del mundo en que se ubiquen sus miembros.

Las redes sociales son una buena manera de mantener el contacto con un grupo amplio de personas o pueden ser útiles para difundir noticias o información a muchas personas al mismo tiempo. Disponen de herramientas para publicar en la red

la información y de manera inmediata está disponible para todos aquellos usuarios que tengan lazos con el que publique dicha información.

Esta capacidad de dar a conocer algo de forma inmediata a cientos de miles de personas es ampliamente utilizada en la actualidad por canales de noticias, diarios y revistas, quienes ven en las redes sociales una forma de mantener informado al público mucho más veloz que con sus propios medios.

II.2 Historia de las redes sociales

Las redes sociales pueden llegar a ser un servicio moderno con escasa trayectoria en la web, debido a que la mayor explosión ha surgido en los últimos años logrando una verdadera masificación en su uso, pero lo cierto es que su origen se remonta a más de una década.

Después de todos estos años, las redes de interacción social se han convertido en uno de los elementos de Internet más difundidos, ofrecen a sus usuarios un lugar común para desarrollar comunicaciones constantes.

Esto es posible, gracias a que los usuarios no sólo pueden utilizar el servicio a través del PC, sino que se ha extendido su uso a dispositivos móviles.

Las primeras redes sociales datan del año 1995, época en la que Internet había logrado convertirse en una herramienta prácticamente masificada. La primera red social se llamaba "Classmates" la cual permitía a las personas de todo el mundo recuperar o mantener el contacto con sus antiguos amigos, ya fueran compañeros de colegio, de la universidad o de distintos ámbitos laborales.

Pero la mayor explosión de las redes sociales ocurrió en el año 2003, año en el cual vieron la luz algunas de las comunidades más populares, que lograron hacer crecer de manera exponencial el uso del servicio. Las más utilizadas en ese momento fueron MySpace, Friendster, Tribe y Xing, entre otras.

En 2004 se lanzó Facebook concebida originalmente como una plataforma para conectar estudiantes universitarios. Su inicio fue en la Universidad de Harvard y durante su primer mes de funcionamiento consiguió que se suscribieran a ella más de 9.500 estudiantes.

En 2006 se inauguró la red Twitter, su definición fue "una corta ráfaga de información". La dinámica de Twitter era simple hay receptores y emisores, los cuales no tienen que establecer lazos de amistad. Los emisores pueden enviar mensajes no superiores a 140 caracteres (tweet) a quienes han elegido seguirlos, mientras los receptores (followers), reciben esos mensajes. Aunque cualquiera puede convertirse en emisor.

Como resumen de los usuarios en 2011 Facebook tenía 600 millones de usuarios repartidos por todo el mundo, MySpace 260 millones, Twitter 190 millones y Friendster apenas 90 millones. Sin embargo el volumen de usuarios en 3 años ha ascendido considerablemente en dos de ellas, a fecha Octubre de 2014, Facebook tenía 1350 millones de usuarios activos en todo el mundo y Twitter en Enero de 2014, tenía 560 millones de usuarios.

Las nuevas redes sociales han dejado obsoletas a la mensajería instantánea, con aplicaciones como Windows Messenger que tuvo su auge con el lanzamiento de la versión 6.0 en el año 2003 que la renovaba por completo añadiendo avatares para el usuario, emoticonos y una interfaz completamente diferente.

Los hábitos sociales han cambiado, en un mundo cada vez más global y multicultural, ha surgido la necesidad de comunicarse con otros de manera rápida, sin ser un impedimento el idioma, la distancia... Este hecho algunos analistas lo ven como

positivo y otros como un punto sin retorno en las relaciones humanas. Las redes sociales han permitido que esta nueva manera de comunicación se expanda, debido a que interconecta a millones de personas.

Las redes sociales son muy populares entre los adolescentes y también han conseguido que los adultos usen sus servicios, es tal la afluencia de usuarios que sitios como Facebook o Google+ son utilizados por las grandes compañías como plataforma publicitaria, incluso relegando la publicidad tradicional en radio, prensa o televisión.

CAPÍTULO III: ESTADO DEL ARTE

En este capítulo se describen las distintas herramientas disponibles en el mercado, así como una breve comparativa de todas, utilizando textos de referencia.

III.1 Introducción

El objetivo del proyecto es crear un programa que identifique automáticamente el idioma de todos los post, para ello se necesita utilizar una herramienta de detección del lenguaje de las disponibles en el mercado.

Tenemos varios tipos de detectores del lenguaje:

- Detectores online:
  - Online: se necesita acceder a una web e introducir el texto manualmente. La información del idioma se obtiene en el momento.
  - Programa descargable: se trata de un software que se instala en el ordenador y tiene un interfaz de usuario. Este permite introducir textos uno a uno y se obtienen los resultados del idioma inmediatamente.
- Detectores programables:
  - Cosiste en un código descargable que se puede integrar en distintas aplicaciones, los lenguajes de programación utilizados son variados (Java, PHP…). Existen múltiples desarrolladores de detectores del lenguaje, unos de código abierto y otros de pago.

Dentro de las distintas herramientas que existen para detectar el lenguaje, los que se han considerado como detectores online son aquellos que el usuario debe introducir los textos manualmente. Por tanto, no es escalable y no cumple con la premisa del proyecto que se trata de realizar el procesamiento del texto de manera remota.

A continuación se muestran los distintos tipos que existen ya que es relevante saber que herramientas existen en el mercado aunque no sean óptimas para este caso.

Se ha considerado relevante mostrar el mismo análisis de los detectores online y los programables para poder comparar los resultados de los detectores online y los detectores programables.

Las comparativas que se realizan posteriormente se hacen con un número muy bajo de muestras, ya que se quiere dar una visión general del funcionamiento.

III.2 Detectores Online

Se han comparado cinco detectores diferentes para ver el comportamiento que presentan según el texto que le introduces.

Las herramientas analizadas son:

- Detect Lang:

  Es una página web versión Beta, que da la opción de detectar el idioma del texto o url introducido, soporta 64 idiomas diferentes y da ejemplos de textos de los idiomas que soporta. La web está traducida a 10 idiomas diferentes para hacerla accesible a muchos usuarios, además anima a los usuarios a traducir la pagina web a otros idiomas en el formato que indican y los propietarios se encargan de actualizarla.

Figura 1: Página Web Detect Lang

- What Language:

  Esta herramienta web da la opción sólo de detectar el idioma del texto introducido, soporta 59 idiomas diferentes. La web está disponible en Inglés y en Japonés.

  Los datos utilizados se recopilan a partir de diferentes versiones lingüísticas de wikipedia, según el autor la wikipedia parece ser una fuente fiable para analizar textos y así poder recopilar datos estadísticos necesarios para identificar el idioma.

Figura 2: Página Web What Language

- Lang Detector:

  Es una página web, que da la opción de detectar el idioma del texto o url introducido, soporta 117 idiomas diferentes y da ejemplos de textos de los idiomas que soporta. La web tiene colgado el API para detectar idiomas de textos y URLs, aunque para poder utilizarlo es necesario comprar la licencia para un año, dependiendo el número de consultas diario el precio varía, el leguaje de programación es PHP.

Figura 3: Página Web Lang Detector

- Detector de idiomas:

  Es una página web, que muestra una gráfica con la probabilidad (en términos relativos) de que la frase introducida esté escrita en alguno de los idiomas soportados (español, latín, catalán, portugués, francés o italiano)

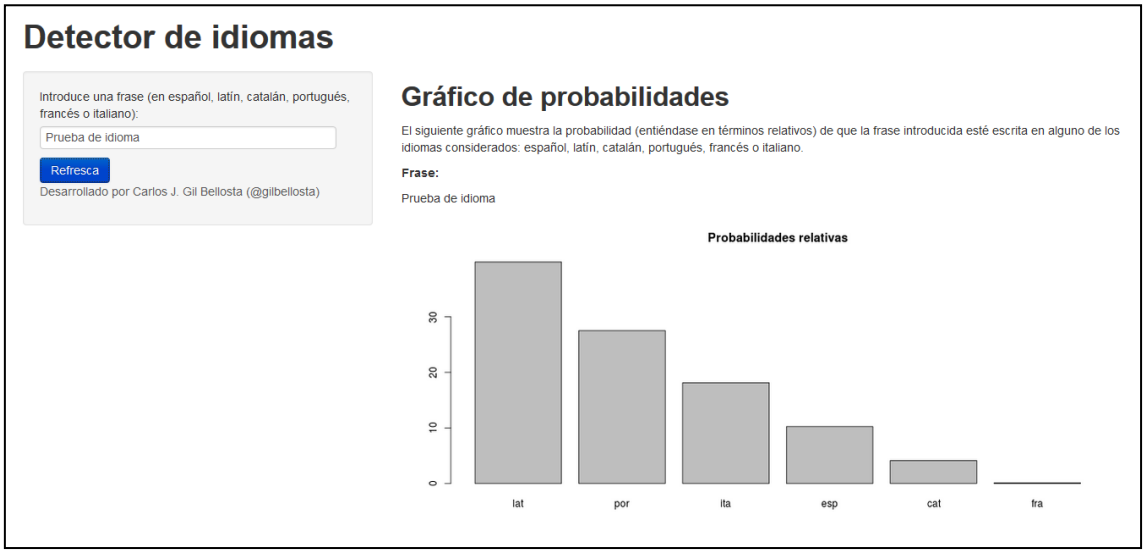


Figura 4: Página Web Detector Idiomas

- Poliglot 3000:
  Es un software disponible para Windows 95/ 98/ NT/ ME/ 2000/ XP/ 2003/ Vista/ 2008/ 7/ 8. Es un identificador automático de idioma que reconoce rápidamente el idioma de cualquier texto, frase o incluso palabras sueltas. Sus principales características es que reconoce más de 400 idiomas, soporta textos Unicode, tiene una cómoda interfaz de usuario y se puede configurar para comprobar sólo para lenguajes populares. Su interfaz multilingüe entre otros soporta inglés, alemán, francés y español.

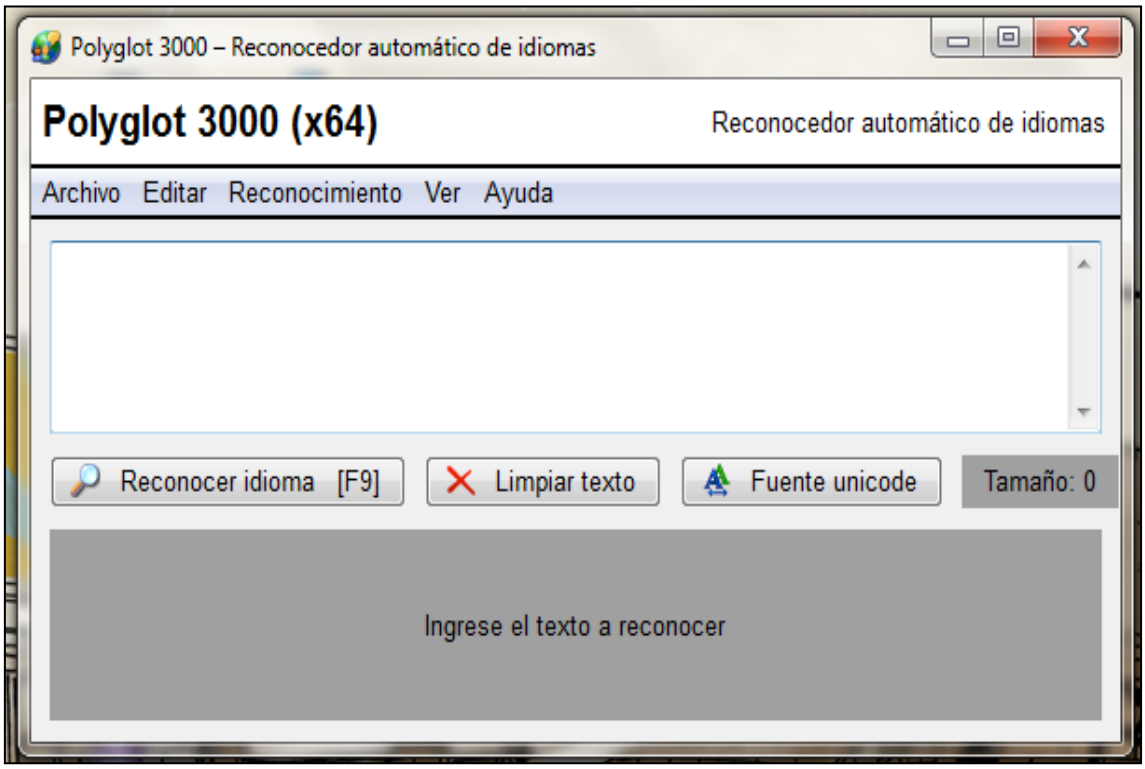


Figura 5: Página Web Detector Idiomas

III.2.1 Descripción del análisis con Detectores Online

La comparativa se ha realizado con cada uno de los detectores online que se han descrito en el apartado anterior y teniendo en cuenta dos factores claves en el proyecto primero el tamaño del texto y el segundo el idioma.

El primer factor es el tamaño del texto, por tanto se han considerado cuatro tipos de textos de distintas longitudes para realizar el análisis:

- Texto largo: Se ha considerado como texto largo una biografía, ya que contiene varias líneas de texto más de 5.
- Texto muy corto: En este caso el titular de un periódico, no consta de más de una línea.
- Texto corto: Se trata del resumen de una noticia de un periódico comprende 2-3 líneas
- Texto muy largo: Se ha considerado un texto muy largo la letra de una canción.

Este factor es importante ya que determina la tasa de acierto en los detectores, con textos muy largos la tasa de acierto tiende a aumentar, sin embargo los textos en los que nos centraremos en capítulos posteriores serán los textos muy cortos.

El segundo factor es el idioma de los textos, se han utilizado el Español, Inglés, Portugués, Alemán, Francés e Italiano. El detector debe ser capaz de identificar bien el idioma independiente del idioma del texto.

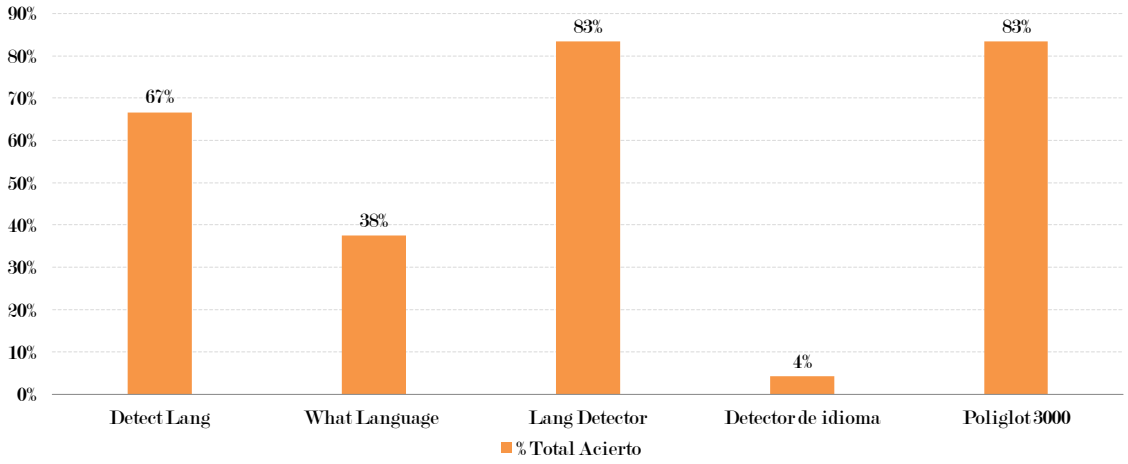
Se ha seleccionado un texto por cada tipo de idioma y por cada tamaño de texto, en el Anexo I se encuentran adjuntos los textos utilizados en el análisis. Conjugando todas las posibilidades se muestra en el apartado siguiente las gráficas con los resultados.

III.2.2 Resultados Análisis Detectores Online

El análisis sólo se ha realizado con un texto de cada tipo e idioma, no se han utilizado más muestras ya que estos detectores a priori no cumplen con los requisitos del proyecto. En los objetivos del proyecto no se pretende examinar si las herramientas existentes dan mejores resultados con textos cortos o largos, pero se ha querido mostrar el análisis previo que se realizó antes con las herramientas existentes en el mercado.

En esta primera gráfica se muestra la tasa de acierto para cada detector online teniendo en cuenta todos los textos y todos los idiomas. Se puede observar que dos de ellos presentan un resultado por encima del 80% y el resto con tasas de acierto muy bajas. Aunque los resultados no son concluyentes debido a que se han tomado pocas muestras, sirven para ver a alto nivel cómo se comporta cada uno. En las próximas

graficas se analizarán las respuestas de los detectores según los tipos de texto y los idiomas seleccionados.

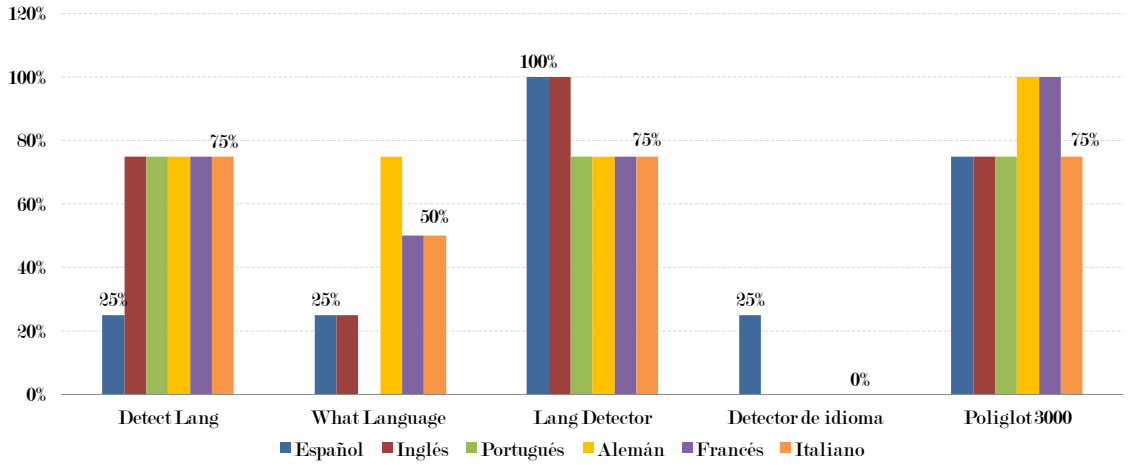


Gráfica 1: % acierto total para cada detector

A continuación, en la gráfica se presenta el % de acierto de cada detector online, por cada lenguaje, es decir, en el eje X se muestra cada tipo de detector y en el eje Y el % de acierto global, teniendo en cuenta todos los textos analizados, en cada columna se representa un idioma diferente.

A diferencia, de la gráfica anterior aquí se desglosa el resultado diferenciando por los idiomas, de los dos que presentaban mejores resultados en la gráfica anterior Lang Detector y Poliglot 3000, no están optimizados para los mismos idiomas, el primero detecta bien en todos los casos el Español y el Inglés, mientras que el segundo el Alemán y el Francés.

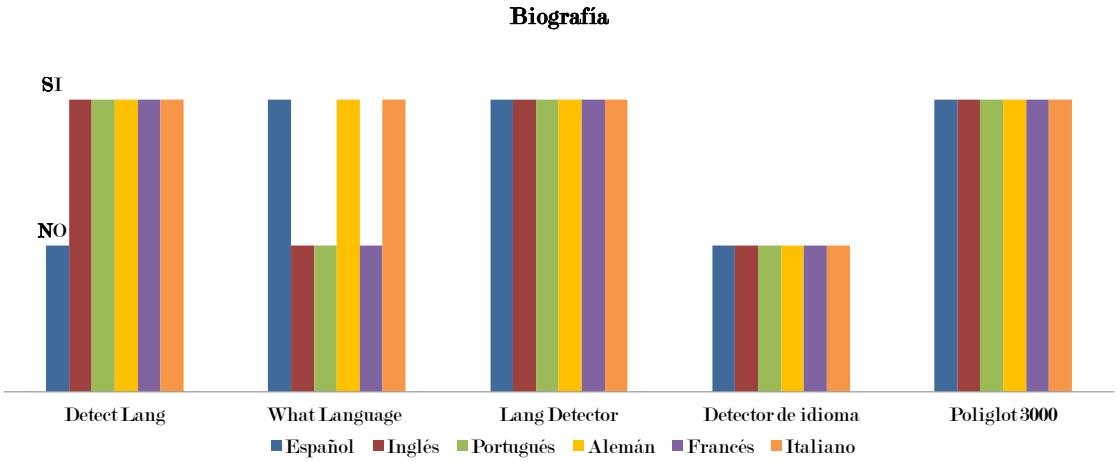
El detector WhatLanguage y el Detector de Idiomas, aunque se mantiene en las gráficas posteriores se observa que la tasa de éxito es muy baja en aquellos idiomas que es capaz de identificar y algunos idiomas, por ejemplo el portugués no puede reconocerlo.



Gráfica 2: % acierto según el idioma para cada detector

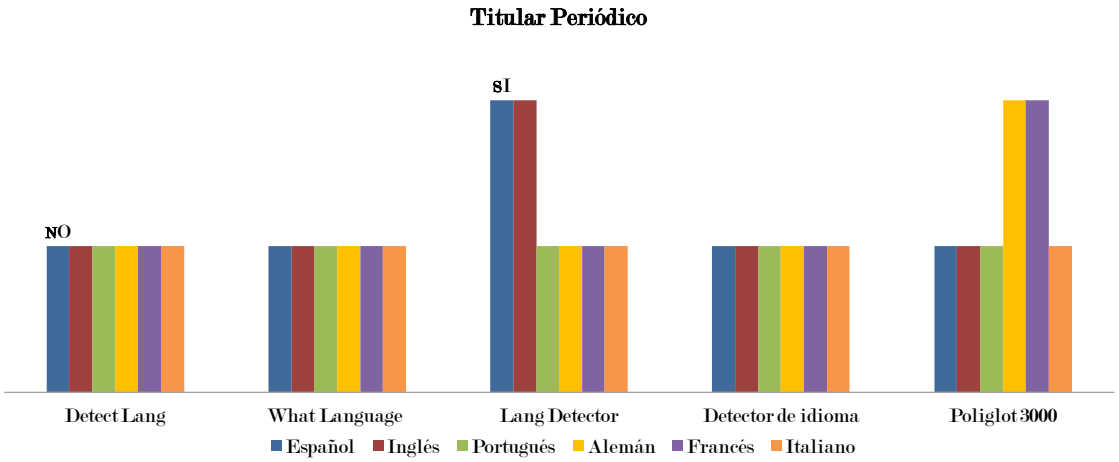
En las próximas cuatro gráficas, se muestra si el texto utilizado de referencia (Biografía, Titular Periódico, Resumen Periódico y Canción) se identifica correctamente en el idioma o no.
Los resultados que se ven en las gráficas son SI, en caso de que el detector identifique correctamente el idioma o NO, en caso contrario.

En la primera de ellas podemos observar que en la mayoría de los casos y de los detectores se identifican todos los idiomas correctamente, algo esperado ya que tiene palabras suficientes para identificar el idioma correctamente.



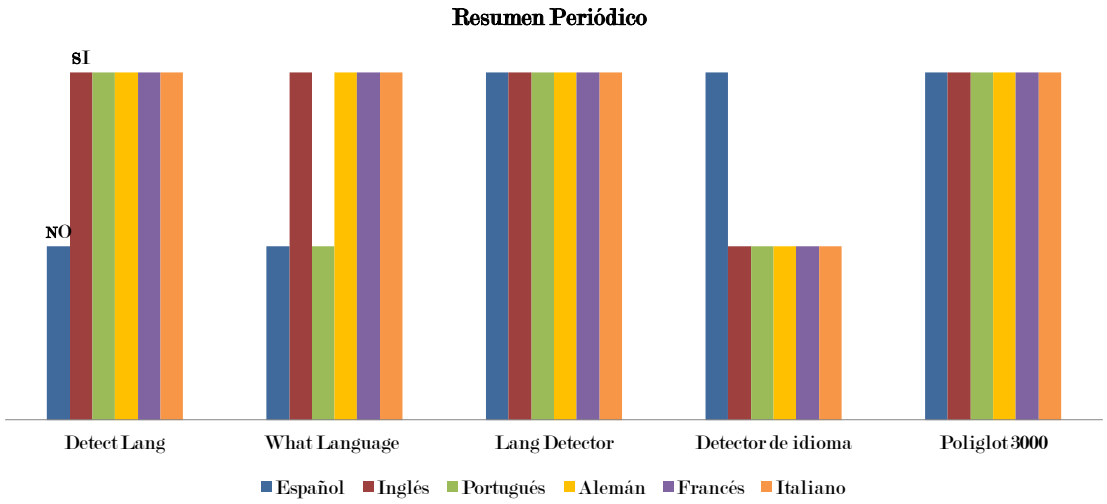
Gráfica 3: Identificación del idioma por los detectores, del texto de referencia Largo (Biografía)

En la segunda gráfica se puede ver el efecto contrario que en la gráfica anterior en la mayoría de los casos y de los detectores no se identifican los idiomas correctamente, la complejidad deriva en que mientras menos palabras la probabilidad de fallo aumenta.
Los idiomas optimizados se mantienen y si reconoce bien en Lang Detector el Español y el Inglés y el Poliglot 3000 el Alemán y el Francés.



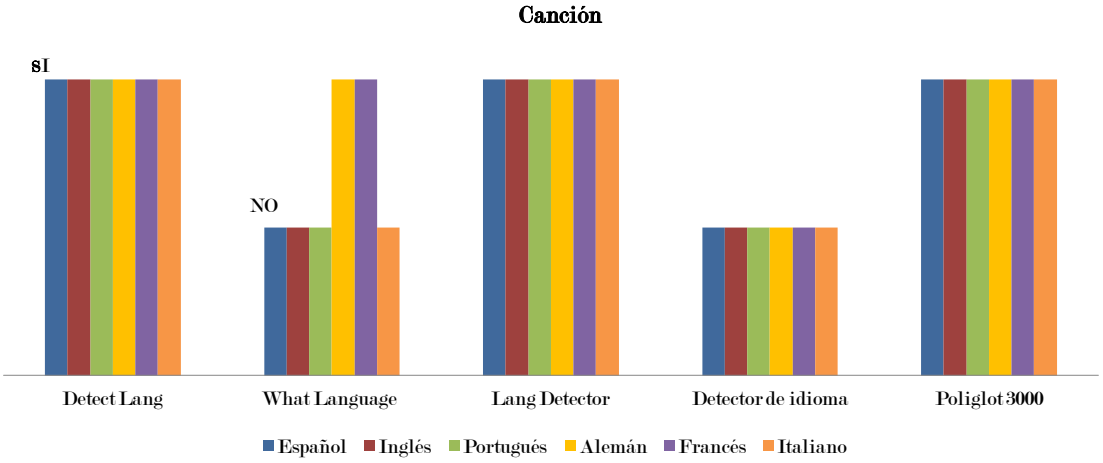
Gráfica 4: Identificación del idioma por los detectores, del texto de referencia Frase (Titular de un periódico)

En la tercera gráfica se puede ver como aumenta significativamente la tasa de aciertos en todos los detectores, el aumento de palabras en el texto aunque no sea largo supone el aumento de la tasa de acierto Lang Detector y el Poliglot 3000 tienen una tasa de éxito del 100% en todos los idiomas.



Gráfica 5: Identificación del idioma por los detectores, del texto de referencia Varias Frases (Resumen de un periódico)

En la última gráfica se observa que Lang Detector y el Poliglot 3000 tienen una tasa de éxito del 100% en todos los idiomas igual que en el caso anterior y se une a ellos Detect Lang. Sin embargo, What Language y Detector de idioma empeoran su tasa de acierto, son muy inestables y la tasa de acierto no es aceptable para ningún idioma.



Gráfica 6: Identificación del idioma por los detectores, del texto de referencia Texto muy largo (Canción)

En resumen, en la Gráfica 2, se puede observar que dentro de los detectores online, utilizando los textos de referencia, Lang Detector y Poliglot 3000 son los que presentan mejores resultados. Estos detectores identifican correctamente todos los idiomas salvo en el texto demasiado corto. Mientras que los detectores What Language y Detector de Idiomas presentan una tasa de acierto muy baja que no es admisible en ningún idioma, ni mejora con los distintos tipos de texto.

Las ventajas de estas soluciones es que son gratuitos, accesibles desde internet o el PC, tienen una interfaz de usuario amigable y permite obtener resultados de manera inmediata.

En este análisis previo al análisis del proyecto, se confirma la teoría que la longitud de los textos influye en los resultados. Debido a que la longitud no será seleccionable ya que la particularidad de los comentarios de Facebook es que los textos son cortos, se buscará la solución óptima para este tipo de textos.

III.3 Detectores Programables

Los detectores programables son aquellos que el código fuente es accesible públicamente y el desarrollador los puede integrar dentro de otro programa.

Existe en el mercado distintas soluciones con el código fuente accesible, que permiten dado un texto identificar el idioma. Estás soluciones están disponibles en varios lenguaje de programación, por ejemplo PHP, C++, Java. Además algunas son gratuitas y otras de pago.

Se han definido dos criterios para la elección de los detectores del lenguaje.
El primero, se necesita que sea accesible a todo el mundo y no tener costes recurrentes por su uso, por ello se han descartado aquellas soluciones del mercado que requerían un pago inicial para obtener el código fuente o un coste recurrente por cada consulta que se realizara. En un entorno social complejo, es clave minimizar cualquier tipo de coste, ya que la herramienta en la que se integrará el detector se pretende utilizar masivamente.
El segundo de los criterios es el lenguaje de programación, debido a que la automatización de los textos se realizará en Java, los detectores de lenguaje que se analizarán serán en el mismo lenguaje. De manera que se minimicen los errores debidos a las incompatibilidades de dos lenguajes de programación distintos y en caso de error el debug sea más sencillo y ágil. Además si se tienen en cuenta futuras evoluciones, se garantiza que teniendo el mismo lenguaje de programación se va a poder evolucionar la herramienta.

Por tanto, de todos los detectores encontrados en Internet, se han seleccionado aquellos gratuitos y que estén desarrollados en Java, quedando la siguiente shortlist para un estudio exhaustivo:

- Detector 1: Langdetec
- Detector 2: GitHub
- Detector 3: Tica

III.3.1 Descripción del análisis con Detectores Programables

La comparativa se ha realizado con cada uno de los detectores programables que se han seleccionado en el apartado anterior con los mismos textos que se han utilizado en el apartado anterior, para poder comparar resultados.

Los datos obtenidos no serán concluyentes para elegir uno u otro, ya que sólo se han utilizado los textos de referencia, pero es posible compararlo con los detectores online al utilizar los mismos textos.

Se han considerado los mismos cuatro tipos de textos de distintas longitudes para realizar el análisis:

- Texto largo: Se ha considerado como texto largo una biografía, ya que contiene varias líneas de texto más de 5.
- Texto muy corto: En este caso el titular de un periódico, no consta de más de una línea.
- Texto corto: Se trata del resumen de una noticia de un periódico comprende 2-3 líneas
- Texto muy largo: Se ha considerado un texto muy largo la letra de una canción.

Es importante la longitud de los textos ya que determina la tasa de acierto de los detectores, con textos muy largos la tasa de acierto tiende a aumentar.
Sin embargo, en el Capítulo siguiente nos centraremos en los textos muy cortos y se realizará un análisis exhaustivo introduciendo en los detectores un volumen elevado de textos cortos, para poder determinar cuál se utilizará en el proyecto.

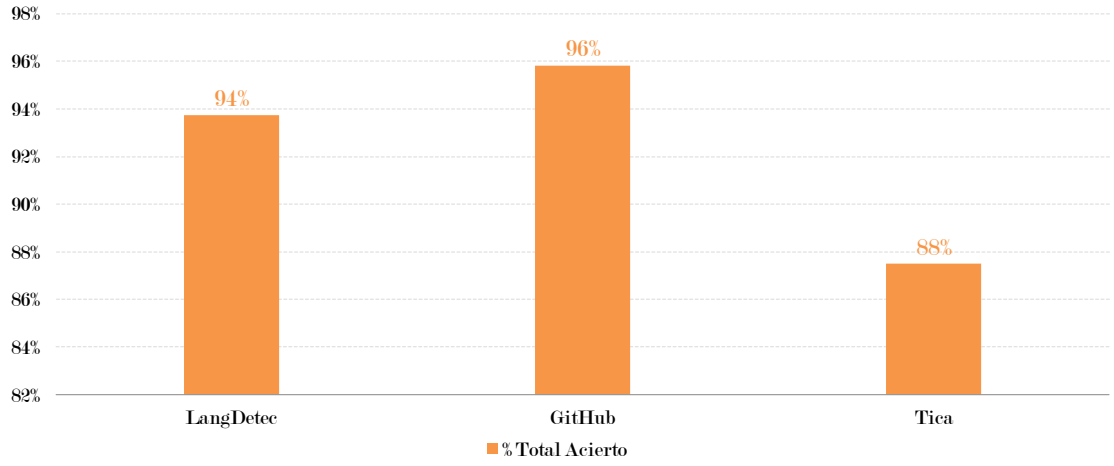
Por otro lado, idioma de los textos que se ha utilizado es el Español, Inglés, Portugués, Alemán, Francés e Italiano. El detector debe ser capaz de identificar bien el idioma independiente del idioma del texto.

Se ha seleccionado un texto por cada tipo de idioma y por cada tamaño de texto, en el Anexo I se encuentran adjuntos los textos utilizados en el análisis. Conjugando todas las posibilidades se muestra en el apartado siguiente las gráficas con los resultados.

III.3.2 Resultados Análisis Detectores Programables

El análisis sólo se ha realizado con un texto de cada tipo e idioma, no se han utilizado más muestras para poder comparar los resultados con los Detectores Online

En esta primera gráfica se muestra la tasa de acierto para cada detector programable teniendo en cuenta todos los textos y todos los idiomas. Se puede observar que todos superan el 80% pero el detector GitHub supera el 95%, resultado muy bueno en comparación con los detectores online que veíamos en el apartado 2.2. Aunque los resultados no son concluyentes debido a que se han tomado pocas muestras y no está enfocado en los textos cortos, se puede decir que al menos la tasa de acierto de los detectores programables para este caso es muy superior a los detectores online.

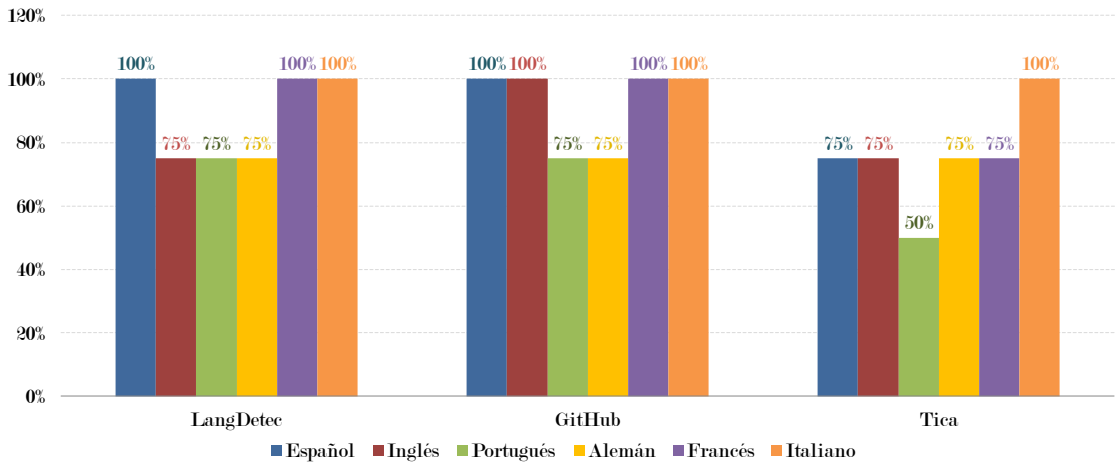


Gráfica 7: % acierto total para cada detector

A continuación, en la gráfica se presenta el % de acierto de cada detector programable, por cada lenguaje, es decir, en el eje X se muestra cada tipo de detector y en el eje Y el porcentaje de acierto global, teniendo en cuenta todos los textos analizados, en cada columna se representa un idioma diferente.

A diferencia, de la gráfica anterior aquí se desglosa el resultado diferenciando por los idiomas. Los tres detectores son capaces de detectar todos los idiomas, aunque no todos con el mismo % de acierto, de los dos que presentaban mejores resultados en la gráfica anterior Detector LangDetec y Detector GitHub, para los idiomas Español, Francés e Italiano los detectan correctamente en el 100% de los casos.

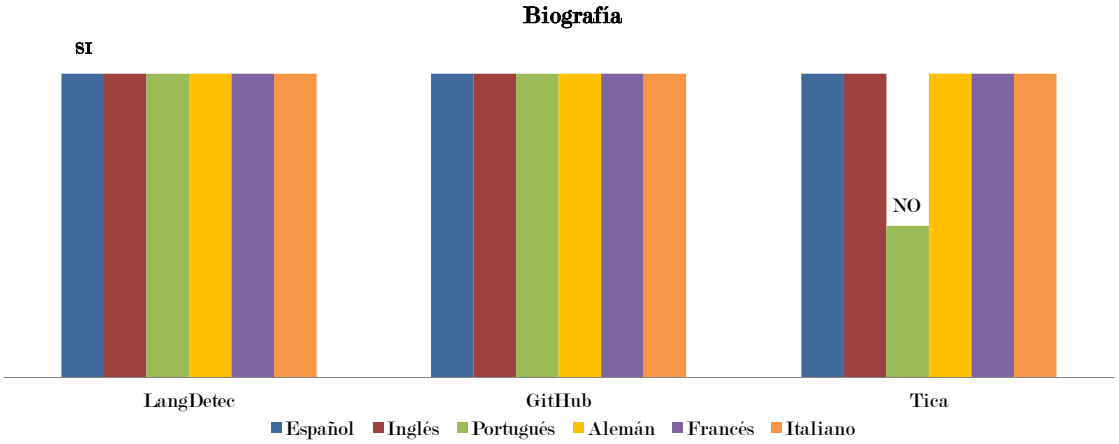
El detector Tica, aunque % de acierto es superior al 80% se observa que la tasa de éxito es muy baja en el portugués.



Gráfica 8: % acierto según el idioma para cada detector

En las próximas cuatro gráficas, se muestra si el texto utilizado de referencia (Biografía, Titular Periódico, Resumen Periódico y Canción) se identifica correctamente en el idioma o no.
Los resultados que se ven en las gráficas son SI, en caso de que el detector identifique correctamente el idioma o NO, en caso contrario.

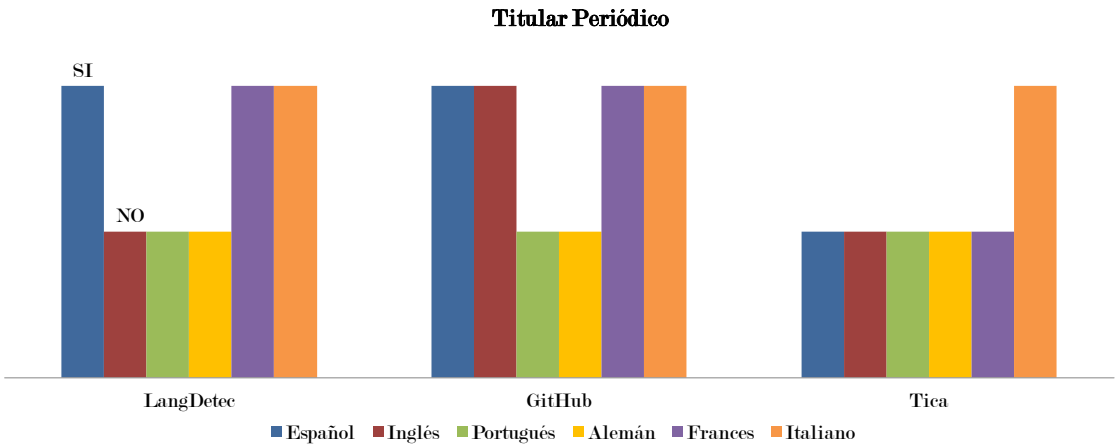
En la primera de ellas podemos observar que en la mayoría de los casos y de los detectores se identifican todos los idiomas correctamente, algo esperado ya que tiene palabras suficientes para identificar el idioma correctamente.



Gráfica 9: Identificación del idioma por los detectores, del texto de referencia Largo (Biografía)

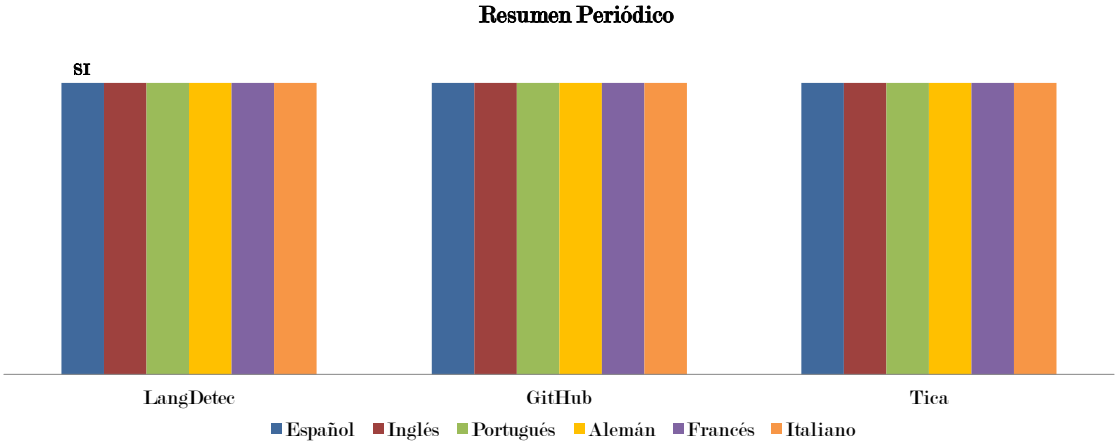
En la segunda gráfica se puede ver el efecto contrario que en la gráfica anterior en la mayoría de los casos y de los detectores no se identifican los idiomas correctamente, la complejidad deriva en que mientras menos palabras la probabilidad de fallo aumenta.

Sin embargo, el Detector GitHub se observa que incluso en textos muy cortos reconoce correctamente el idioma, a excepción del Portugués y del Alemán. Estos buenos resultados del Detector GitHub para textos muy cortos se analizarán con más detalle en el siguiente capítulo y se comprobará si se mantienen estos resultados.



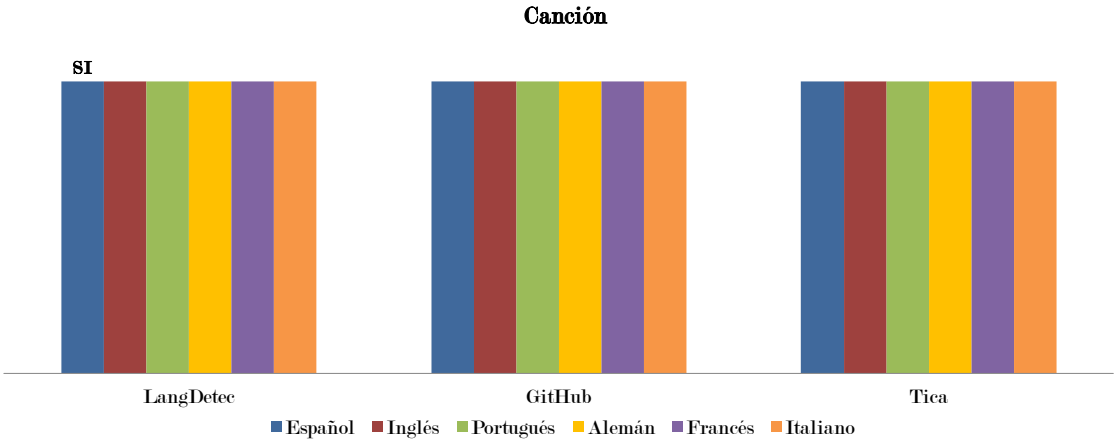
Gráfica 10: Identificación del idioma por los detectores, del texto de referencia Frase (Titular de un periódico)

En la tercera gráfica se puede ver como aumenta significativamente la tasa de aciertos en todos los detectores, el aumento de palabras en el texto aunque no sea largo supone el aumento de la tasa de acierto, en los 3 detectores se identifican correctamente todos los idiomas.



Gráfica 11: Identificación del idioma por los detectores, del texto de referencia Varias Frases (Resumen de un periódico)

En la última gráfica se observa el mismo resultado que en el resumen del periódico los tres detectores identifican correctamente todos los idiomas, a diferencia de los detectores online para la canción sí que se reconocen correctamente todos los idiomas y se mantiene la tasa de acierto.



Gráfica 12: Identificación del idioma por los detectores, del texto de referencia Texto muy largo (Canción)

En resumen, en la Gráfica 8, de los detectores programables, utilizando los textos de referencia, el Detector GitHub es el que presenta mejores resultados, pero con tan pocas muestras no podemos tener ningún resultado concluyente. Estos detectores identifican correctamente todos los idiomas salvo en el texto demasiado corto, será necesario una batería de pruebas más extensa para tener un porcentaje de acierto representativo.

Las ventajas de estas soluciones es que son gratuitos y permiten integrarlos dentro de otros programas.

En este análisis previo al análisis del proyecto, se confirma la teoría que la longitud de los textos influye en los resultados. Debido a que la longitud no será seleccionable ya que la particularidad de los comentarios de Facebook es que los textos son cortos, se buscará la solución óptima para este tipo de textos y se concluirá si el Detector GitHub es el más óptimo para el proyecto o por el contrario otro presenta mejores resultados.

CAPÍTULO IV: EVALUACIÓN RENDIMIENTO

En este Capítulo se va a evaluar el rendimiento de los 3 detectores programables elegidos anteriormente, esto nos permitirá seleccionar el detector de texto que se integrará en la herramienta y por tanto se utilizará en el caso práctico.

IV.1 Descripción del escenario

Se pretende evaluar el performance de los tres detectores programables, en este caso el análisis se centrará en textos cortos ya que es el objetivo de la herramienta estar optimizada para identificar correctamente textos cortos, concretamente post de las redes sociales.

Para el análisis se dispone de un conjunto de post de Facebook de YouTube y Starbucks proporcionado por el Institut Mines-Telecom SudParis con el que colabora mi tutor.
Se han realizado tres test, para el primero se ha seleccionado de manera aleatoria una muestra representativa de mensajes, concretamente 500 para introducirlos en la herramienta con cada detector y posteriormente se verificará manualmente el resultado obtenido. Para el segundo de los test se utilizarán todos los mensajes proporcionados, 498.499 en total. Estos textos los procesará cada uno de los detectores y se obtendrán conclusiones. Y en el tercer test, se mide el tiempo de procesado de cada uno de los detectores, para ello se ha utilizado una muestra aleatoria de 214.731 comentarios de los usuarios con los cuales se medirá el tiempo que tarda cada herramienta en determinar el idioma de cada comentario.

Se compararán los 3 detectores del lenguaje seleccionados en el capítulo anterior, cada uno de ellos según su desarrollador es capaz de identificar un número distinto de idiomas, concretamente:

- El primero de los detectores de texto es Langdetec, detecta 53 idiomas diferentes, tiene la problemática que da un java null point exception cuando no se le introducen letras, por tanto es necesario realizar un procesado previo de todos los mensajes y descartar aquellos que solo tienen números o caracteres especiales.
- El segundo de los detectores de texto es GitHub, detecta 21 idiomas diferentes, este permite introducir cualquier texto independiente de si son sólo letras o caracteres especiales, no es necesario realizar un tratamiento previo de los mensajes.
- El tercero de los detectores de texto es Tica, detecta 27 idiomas diferentes, este permite introducir cualquier texto independiente de si son sólo letras o

caracteres especiales, no es necesario realizar un tratamiento previo de los mensajes.

IV.2 Descripción de la herramienta

La herramienta que automatiza los textos e integra el detector del lenguaje tiene la misma arquitectura para los tres detectores. Se ha realizado una arquitectura común compatible con los tres detectores del lenguaje. Esta herramienta es un programa en Java que se compone de los siguientes métodos:

- Main: es el método principal desde el cual se ejecuta el programa, tiene todas las variables globales y necesarias para la correcta ejecución.
- Leer: se encarga de leer la estructura de ficheros de los post en la ruta indicada, así como obtener dentro de cada post los mensajes que han escrito los usuarios en la red social.
- Escribir: método que guarda en fichero (en la ruta de salida) la siguiente información, nombre de carpeta, nombre del post, mensaje del usuario y el idioma del mensaje. Con el fichero generado de salida se realiza el análisis de los resultados.
- Identificar Idioma: este método es el encargado de integrar el detector del lenguaje en el programa y es el método que varía según uno u otro ya que la forma de invocarlo varía según lo haya desarrollado el propietario.

El programa se ejecutará utilizando un PC y la máquina virtual de Java.

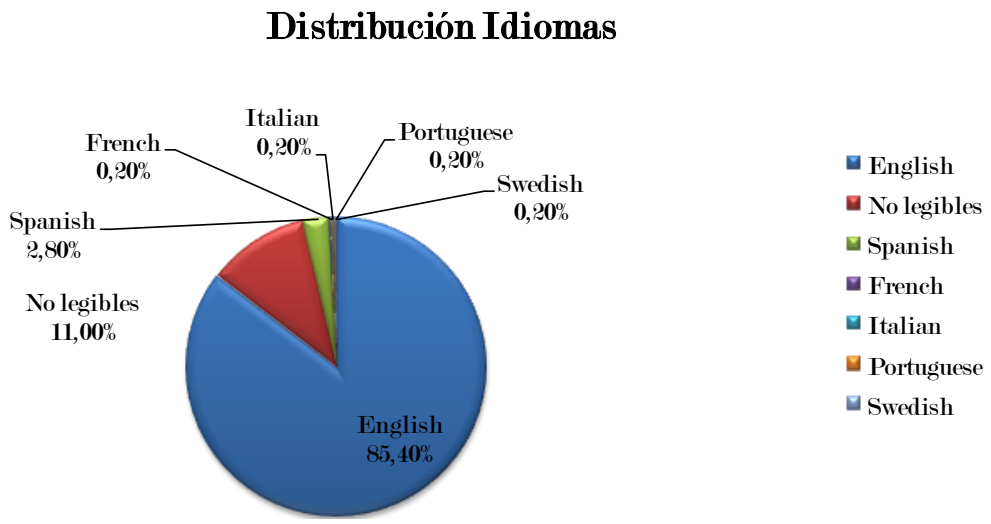
IV.3 Resultados del análisis

En esta sección se muestran los resultados de los análisis que se han realizado con los tres detectores, el objetivo es determinar cuál es el óptimo para nuestro proyecto.

IV.3.1 Clasificación manual de los Idiomas

El primer test que se ha realizado ha sido introducir los 500 mensajes en la herramienta, con cada uno de los detectores. Se ha obtenido el idioma para cada mensaje y manualmente se le ha asignado el idioma real a cada mensaje.

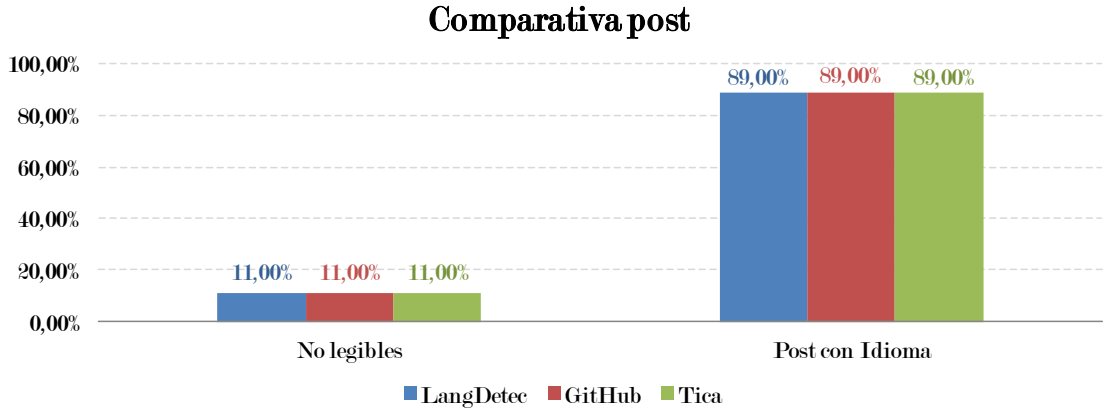
En el siguiente gráfico se muestran los idiomas reales de los 500 mensajes, se puede observar que el mayor porcentaje, un 85,4% está escrito en Inglés (427 mensajes).



Gráfica 13: Distribución real del idioma de los 500 mensajes

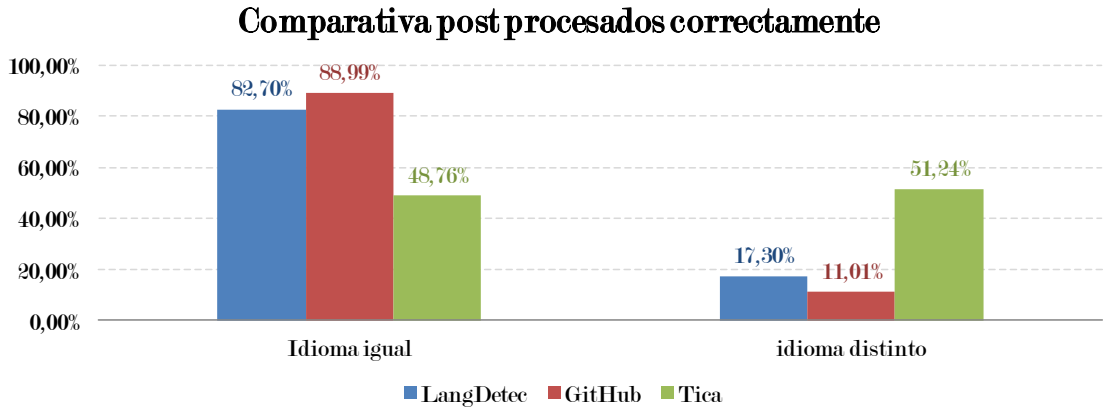
En este primer análisis se puede observar que tenemos un 11% (55 mensajes) que no son válidos para medir el performance de los detectores. Las redes sociales se caracterizan por usar un lenguaje coloquial con onomatopeyas y caracteres especiales del tipo =), :O, :P, Yeehhh, yujuu, que no es posible asignarle un idioma ya que son universales.

Esta peculiaridad del lenguaje de las redes sociales vemos que introduce un 11% de error en la herramienta que se deberá tener en cuenta en el caso práctico.



Gráfica 14: Comparativa de los 3 detectores con 500 los mensajes, mensajes con idioma vs mensajes no legibles

El segundo análisis se ha realizado sólo con aquellos mensajes que tienen asignado un idioma por los detectores del lenguaje, es decir 445 mensajes.
Manualmente se ha comparado el idioma asignado por los detectores con el idioma real, dando lugar a la siguiente gráfica.
En ella se puede observar que el detector que tiene la mejor tasa de éxito es el detector GitHub, aunque el detector LangDetec también un porcentaje de éxito alto. El detector Tica, tiene una tasa de acierto similar a la tasa de fallo, algo no aceptable para la herramienta.

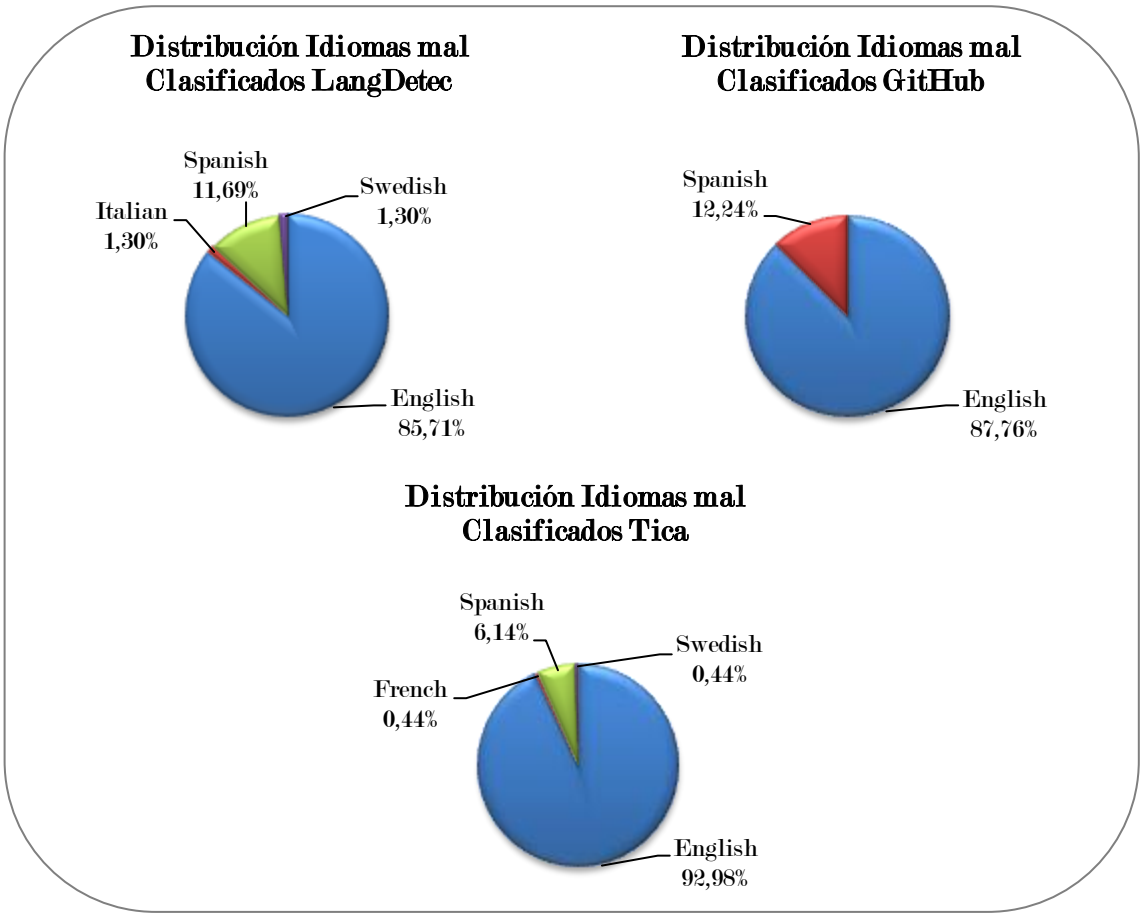


Gráfica 15: Gráfica mensajes procesados correctamente vs fallidos por cada detector

En las siguientes gráficas se muestra la distribución de los idiomas que no reconocen bien cada detector. En los tres detectores se observa que el Inglés es el principal idioma que no reconoce bien, seguido del Español. Esto tiene sentido ya que la mayoría de los mensajes analizados están en inglés y por tanto la probabilidad de que falle en este idioma aumenta.

En el detector LangDetec y GitHub el comportamiento es similar, sin embargo el detector Tica, la tasa de fallo supera el 50%, una de las posibles causas de que este detector tenga tan alta la tasa de fallo es que no está optimizado para el inglés, ya que de los 228 mensajes que no se ha identificado bien el idioma el 92,98% (212 mensajes) eran en inglés.

En las gráficas se observa la distribución de los idiomas mal clasificados por cada detector.



Gráfica 16: Distribución por idioma de los mensajes mal clasificados por cada detector

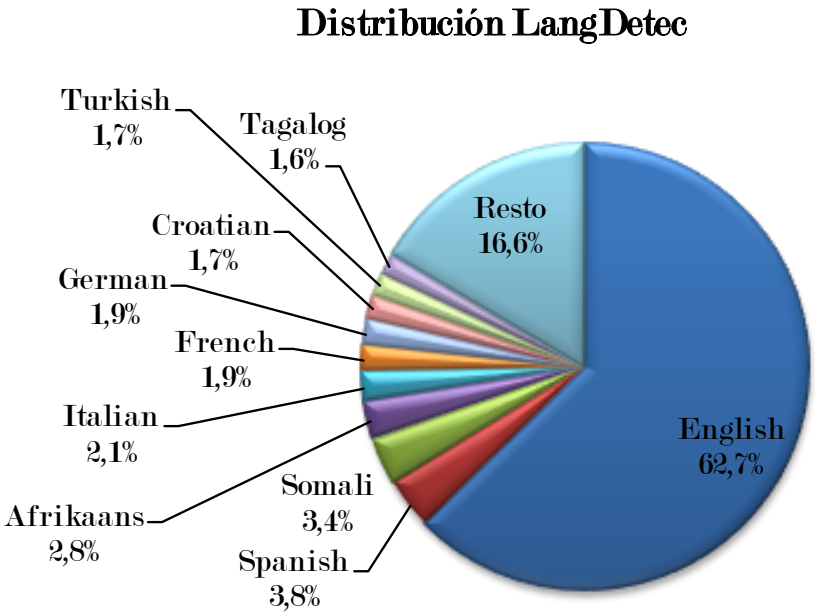
IV.3.2 Clasificación automática de los Idiomas

En este segundo test se pretende clasificar de manera automática el idioma de los comentarios con cada detector, se dispone de una muestra mayor de comentarios (un total de 498.499 mensajes) de los usuarios de las páginas de Facebook de YouTube y Starbucks.

Se van a extrapolar los resultados obtenidos manualmente a una muestra más amplia, teniendo en cuenta las tasas de fallo obtenidas anteriormente y el problema que presenta el detector Tica para identificar los textos en Inglés.

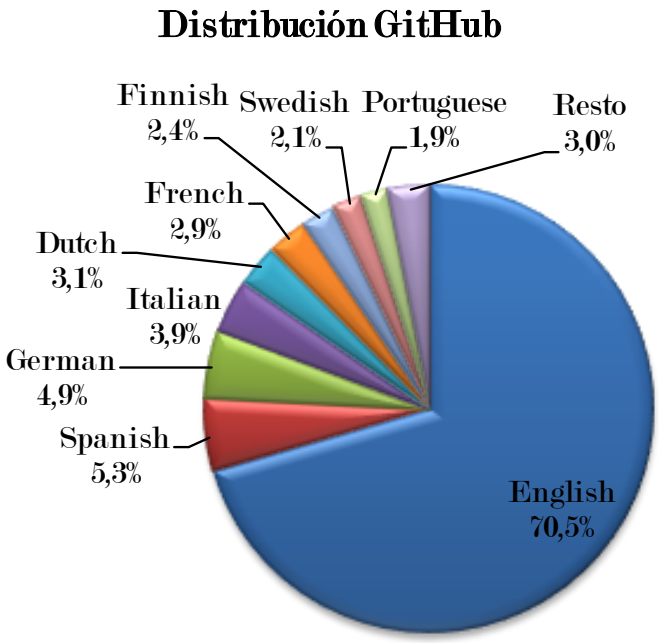
Se ha introducido todos los mensajes en cada detector y este ha clasificado cada mensaje con su idioma, en las siguientes gráficas se muestran la distribución en porcentaje de los idiomas de los mensajes por detector, aquellos idiomas que el porcentaje de representación es inferior al 1,5% se han agrupado bajo la etiqueta de Resto.

El Detector LangDetec, el 62,7% de los mensajes ha determinado que el idioma es el inglés, el 48,3% está distribuido en los diferentes idiomas. Dentro de la etiqueta de Resto se agrupan 20 idiomas más, los cuales su suma supone el 16,6%.



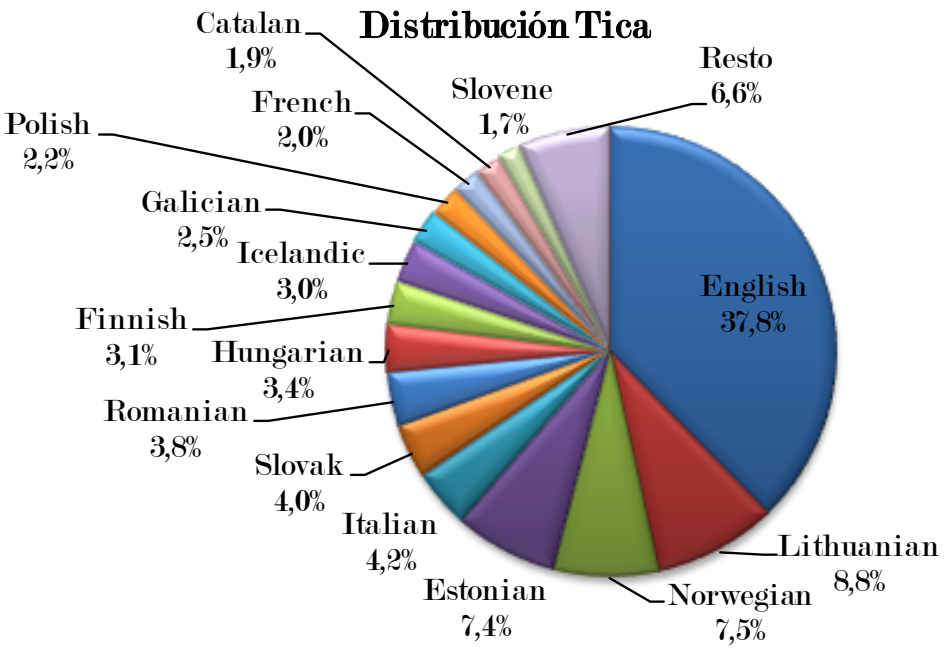
Gráfica 17: Distribución por idioma, clasificado por el detector LangDetec

El Detector GitHub, el 70,5% de los mensajes ha determinado que el idioma es el inglés, el 29,5% está distribuido en los diferentes idiomas. Dentro de la etiqueta de Resto se agrupan 14 idiomas más, los cuales su suma supone el 3%.



Gráfica 18: Distribución por idioma, clasificado por el detector GitHub

El Detector Tica, el 37,8% de los mensajes ha determinado que el idioma es el inglés, el 62,2% está distribuido en los diferentes idiomas. Dentro de la etiqueta de Resto se agrupan 10 idiomas más, los cuales su suma supone el 6,6%, el Español se encuentra dentro de la etiqueta Resto y solo supone el 1,3% de los mensajes.



Gráfica 19: Distribución por idioma, clasificado por el detector Tica

Se puede observar que en los tres detectores, el idioma predominante es el Inglés, esto es coherente con la muestra aleatoria de 500 mensajes del primer test, se observaba que el Inglés era el idioma más predominante.

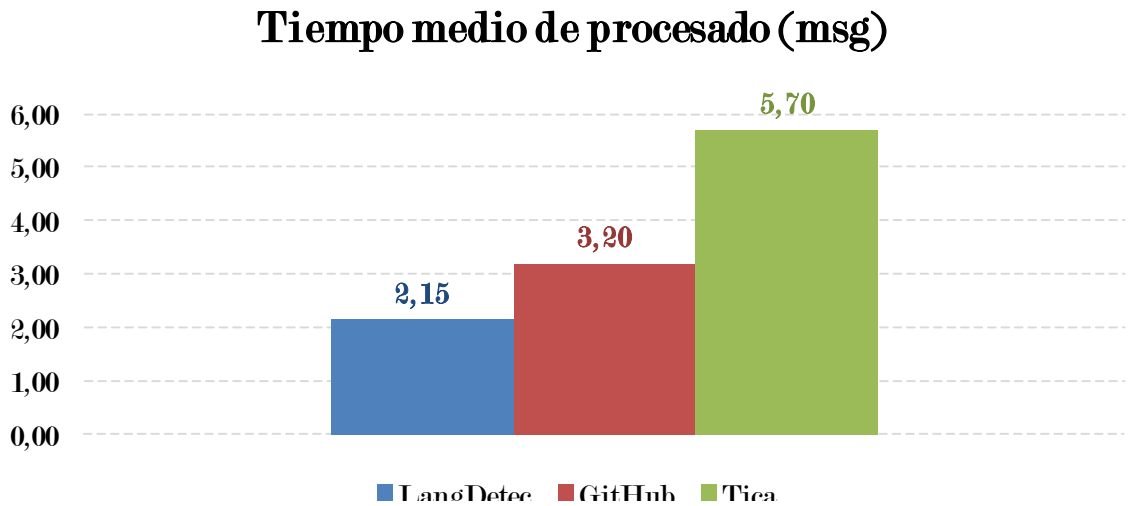
Es destacable que en el detector Tica el porcentaje del Inglés, se reduce hasta el 37,8%, casi la mitad que en el resto de detectores, esto se debe a que la tasa de fallo es aproximadamente el 50% como se ha determinado en el test anterior y aparecen multitud de idiomas que no corresponden con los idiomas identificados en la muestra de 500, ni en los otros detectores.

IV.3.3 Evaluación del tiempo de procesado

En el tercer test se ha evaluado el tiempo que tarda cada herramienta en procesar los comentarios, se ha tomado una muestra aleatoria de 214.731 comentarios de los usuarios de las páginas de Facebook de YouTube y Starbucks.

El tiempo se ha medido tomando el tiempo inicial y el tiempo final de cada mensaje procesado, la diferencia es el tiempo que se considera en procesar cada mensaje.

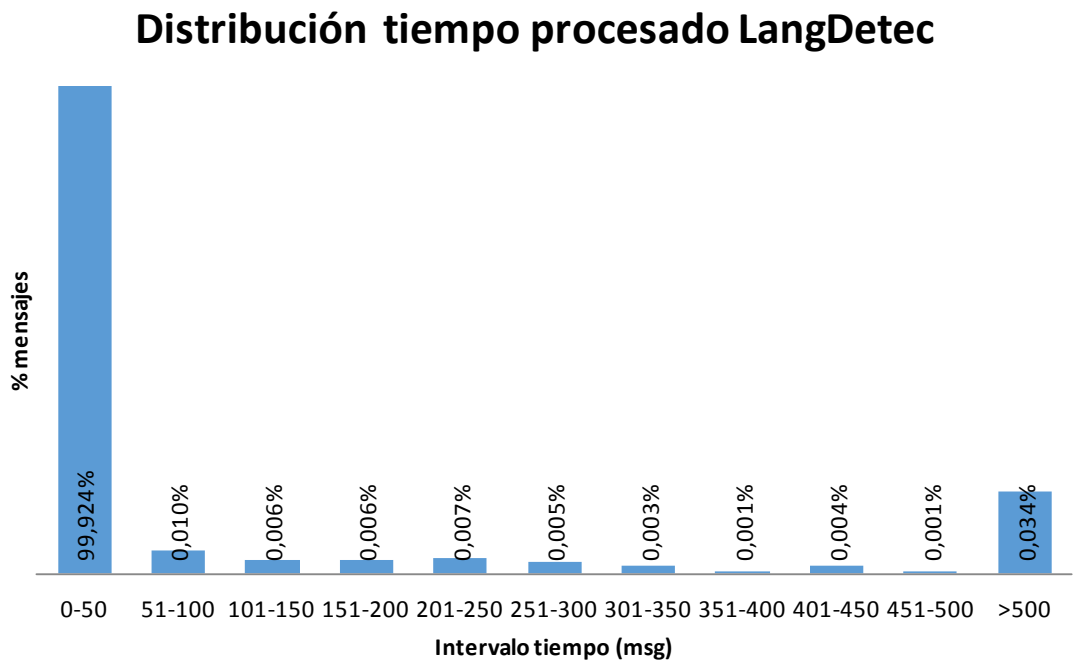
En esta gráfica se muestra el tiempo promedio que tarda la herramienta con cada detector en procesar todos los mensajes. Se puede ver que el detector LangDetec y GitHub el tiempo promedio de procesado de los mensajes es similar, sin embargo el detector Tica tiene un tiempo medio de procesado que duplica a los otros dos.



Gráfica 20: Tiempo de medio procesamiento de cada detector de todos los mensajes

En la siguiente gráfica se muestra el tiempo de procesado de todas las muestras distribuidas por intervalos de tiempo para el detector LangDetec, en el eje vertical se puede observar el porcentaje de mensajes y en el eje horizontal el intervalo de tiempo de procesado en milisegundos en el que se encuentra.

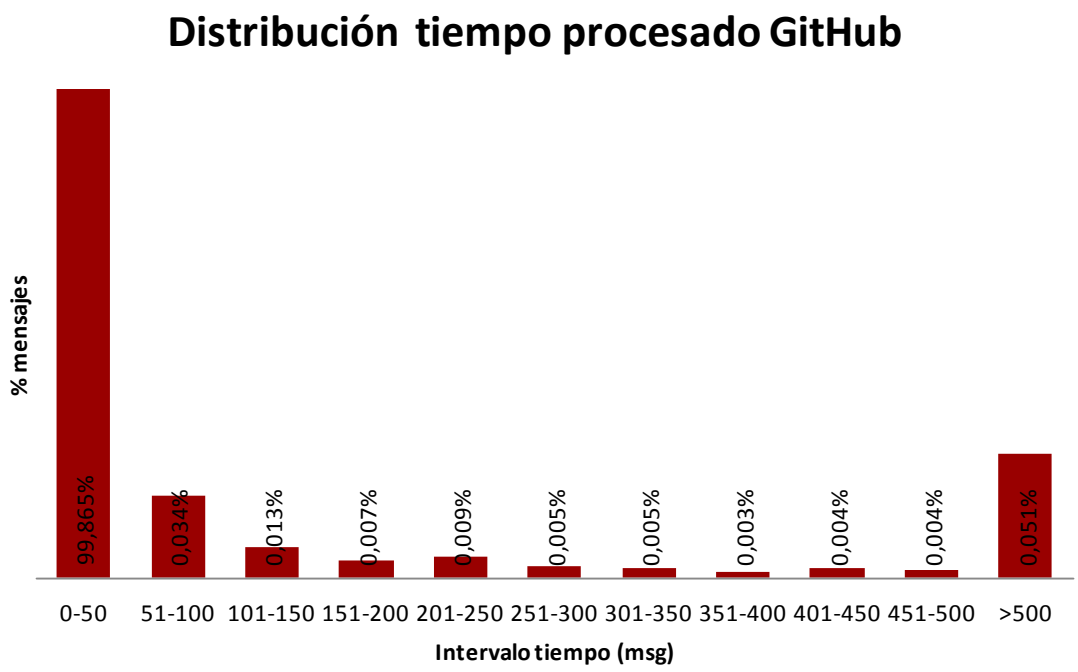
Se puede observar que el mayor número de muestras se concentra en el primer intervalo, esto es un valor muy bueno ya se puede procesar un volumen muy elevado en un tiempo reducido.



Gráfica 21: Distribución del tiempo de procesado del detector LangDetec

En esta gráfica se muestran los resultados del tiempo de procesado de todas las muestras del detector GitHub, en el eje vertical se puede observar el porcentaje de mensajes y en el eje horizontal el intervalo de tiempo de procesado en milisegundos en el que se encuentra.

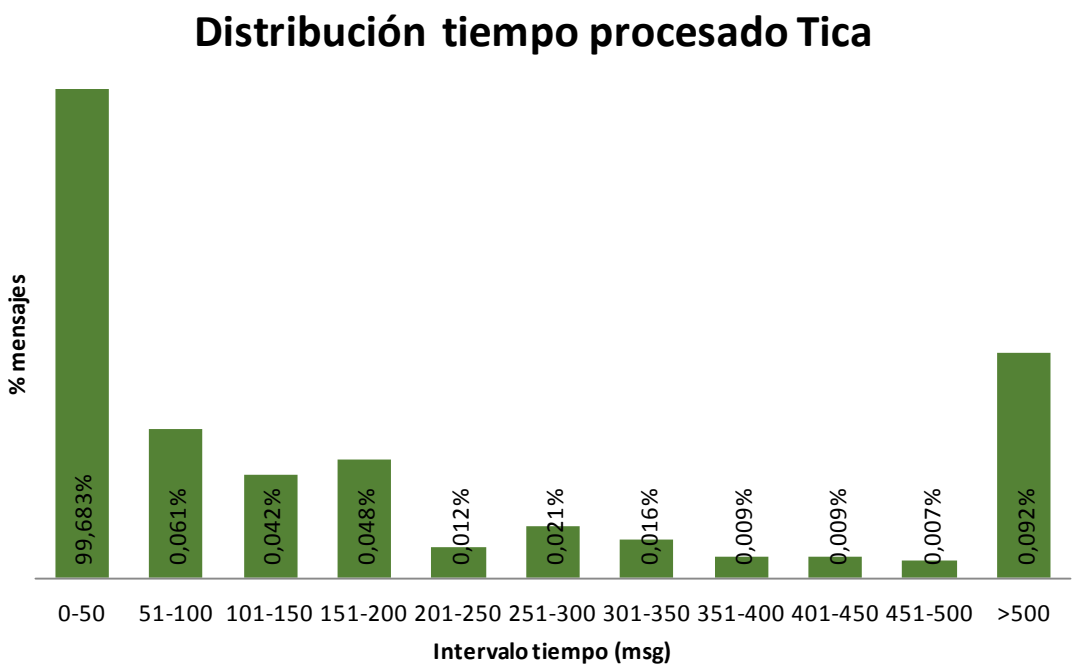
La distribución igual que en el detector LangDetec también concentra cerca del intervalo 0-50, en este caso el 99,86% de los mensajes, siendo coherente con la primera gráfica en la que se veía que el detector LangDetec y GitHub tenían un comportamiento similar.



Gráfica 22: Distribución del tiempo de procesado del detector GitHub

En esta gráfica se muestran los resultados del tiempo de procesado de los comentarios del detector Tica, en el eje vertical se puede observar el porcentaje de mensajes y en el eje horizontal el intervalo de tiempo de procesado en milisegundos en el que se encuentra.

La distribución también presenta un volumen elevado cercano a o pero un mayor número de muestras distribuidas en el eje horizontal, lo que supone que se eleve el tiempo medio de procesado.



Gráfica 23: Distribución del tiempo de procesado del detector Tica

IV.3.4 Resumen de los resultados obtenidos

Según los test realizados, las prestaciones que ofrece la herramienta integrando cada uno de los detectores se resumen a continuación:

Detector Lang Detect:

- De los 500 comentarios seleccionados para el análisis 55 (11%) no han podido ser evaluados por ser palabras ininteligibles. Por tanto con este detector suponemos un 11% de error que se deriva de esta causa.
- De los comentarios evaluados 445 (89%), detecta correctamente el idioma en un 82,47% de los casos.
- El porcentaje de error (los que detecta incorrectamente y los que no se pueden analizar) es de 25,2%.
- El tiempo medio en procesar 214.731 mensajes es de 2,15 milisegundos.

Detector GitHub:

- De los 500 comentarios seleccionados para el análisis 55 (11%) no han podido ser evaluados por ser palabras ininteligibles. Por tanto con este detector suponemos un 11% de error que se deriva de esta causa.
- De los comentarios evaluados 445 (89%), detecta correctamente el idioma en un 88,99% de los casos.
- El porcentaje de error (los que detecta incorrectamente y los que no se pueden analizar) es de 21,00%.
- El tiempo medio en procesar 214.731 mensajes es de 3,20 milisegundos.

Detector Tica:

- De los 500 comentarios seleccionados para el análisis 55 (11%) no han podido ser evaluados por ser palabras ininteligibles. Por tanto con este detector suponemos un 11% de error que se deriva de esta causa.
- De los comentarios evaluados 445 (89%), detecta correctamente el idioma en un 48,76% de los casos.
- El porcentaje de error (los que detecta incorrectamente y los que no se pueden analizar) es de 56,40%.
- El tiempo medio en procesar 214.731 mensajes es de 5,70 milisegundos.

IV.4 Conclusiones

Con los análisis realizados, es necesario tomar la decisión de cuál de los 3 detectores del lenguaje será el elegido para integrarse en la herramienta que permitirá evaluar el efecto de las noticias, promociones, campañas… en las redes sociales.

Los resultados obtenidos en el detector Tica son inaceptables, debido a que la tasa de éxito es inferior al 50% y el tiempo de procesado de los mensajes es el más elevado. Por este motivo se descarta esta solución.

Es necesario llegar a un compromiso entre tiempo de procesado y tasa de acierto. Concretamente para este proyecto es muy importante que la tasa de éxito para textos cortos sea lo más alta posible, aunque esto suponga un incremento en el tiempo de procesado.

De los otros dos aunque el detector LangDetec es más rápido y el número de idiomas soportado es mayor, la tasa de éxito es menor que en el detector GitHub. Dentro de todos los factores el que más peso tiene es la tasa de éxito ya que es clave que los resultados sean fiables y que los textos no necesiten un procesado previo.

Por tanto el analizador elegido para realizar el caso práctico es el detector GitHub.

CAPÍTULO V: CASO PRÁCTICO

En este capítulo se muestran unos ejemplos prácticos en los que se podría utilizar la herramienta. Se va a utilizar una empresa activa en las redes sociales, que promociona sus productos a través de ellas.

V.1. Introducción

El objetivo del caso práctico es mostrar el potencial y la utilidad de la herramienta en un caso real. Para ello, se va a utilizar la empresa King Digital Entertaiment, es una compañía que se dedica a desarrollar juegos en web, en móviles (IOS, Android y Windows Phone) y en Facebook.

La empresa se creó en 2003 entre los fundadores se encuentran Riccardo Zacconi director ejecutivo y Melvyn Morris CEO de la compañía. La empresa cuenta con más de 600 empleados, entre empleados propios y autónomos que colaboran para la puesta en producción.

King Digital Entertaiment, se caracteriza por ofrecer a sus usuarios juegos que permiten guardar el progreso independientemente del dispositivo desde el que jueguen, esto es debido a que a través de la cuenta de Facebook se sincronizan los juegos sin perder los usuarios el progreso de sus juegos.

De los múltiples juegos que tiene la empresa disponible, nos vamos a centrar en 8 de ellos Pet Rescue Saga, Pyramid Solitaire Saga, Candy Crush Saga, Bubble Witch 2 Saga, Farm Heroes Saga, Diamond Digger Saga, Papa Pear Saga y Pepper Panic Saga.

Pet Rescue Saga

Se lanzó en Q4 del 2012 para Facebook y en Q2 del 2013 para dispositivos móviles. Es un puzle basado en bloques, en el que hay que ir destruyendo conjuntos de ladrillos del mismo color para poder liberar a varios animales. Cuando las piezas caen, los animalitos desaparecen y se consigue el objetivo.

Pyramid Solitaire

Se lanzó en Facebook en Mayo de 2012 y en dispositivos móviles en Julio 2014. Es un juego de cartas, concretamente un solitario ambientado en Egipto que ofrece una gran cantidad y variedad de modos de juego que ponen a prueba la habilidad del jugador para ordenar, coleccionar y emparejar cartas.

Candy Crush

Es un videojuego para terminales móviles y Facebook que se lanzó el 12 de abril de 2012 en Facebook y cuya aplicación para plataformas móviles vio la luz el 14

de noviembre del mismo año. Desde marzo de 2013, "Candy Crush Saga" superó a FarmVille 2 como el juego más popular en Facebook, con 45.6 millones usuarios promedio al mes. Actualmente cuenta con 150 millones usuarios únicos mensuales y 54 millones de personas juegan a diario, siendo además la aplicación nº1 en ingresos en las tiendas Google Play y App Store y una de las páginas de Facebook más populares, con 75 986 554 de "me gusta".

Bubble Witch 2

Se lanzó en Mayo del 2014 en plataformas móviles y en Facebook, es una evolución del antiguo Bubble Witch Saga, cuyo objetivo principal consistía en disparar desde un cañón inferior a las bolas que van apareciendo en la pantalla para eliminarlas. Esta nueva versión se lanzó con una imagen renovada, más modos de juego y otras mejoras.

Farm Heroes

Se lanzó en Q2 del 2013 en Facebook y en Enero del 2014 para dispositivos móviles. El juego se trata de un nuevo y mejorado club de granja, en el cual los usuarios tienen que intercambiar y combinar cosechas e impedir que el Mapache Maquiavélico destroce los hermosos terrenos de cultivo.

Diamond Digger

Se lanzó en Junio 2014 en Facebook y en Agosto del 2014 en dispositivos móviles. Los usuarios tienen que explorar, explotar y excavar su propio camino a través de decenas de deslumbrantes mundos subterráneos desenterrando una gran variedad de joyas de colores y tesoros especiales.

Papa Pear

Se lanzó en 2013 en cambio, rompe este esquema sencillo que tantos fans está captando para los juegos de King. Desarrollado en el estudio de Barcelona, es más parecido a Angry Birds que a Candy Crush. Se tiene que conseguir disparar bolitas hacia objetivos concretos pasando por dianas que hacen ganar puntos.

Pepper Panic

Se lanzó en Q4 del 2013 para Facebook. Se trata de pasar los niveles intercambiando frutas y verduras de sus posiciones para formar combinaciones de tres o más del mismo tipo. Para poder avanzar a los siguientes niveles se tendrá que completar el objetivo de cada uno e ir ganando experiencia para descubrir nuevos elementos y nuevas combinaciones.

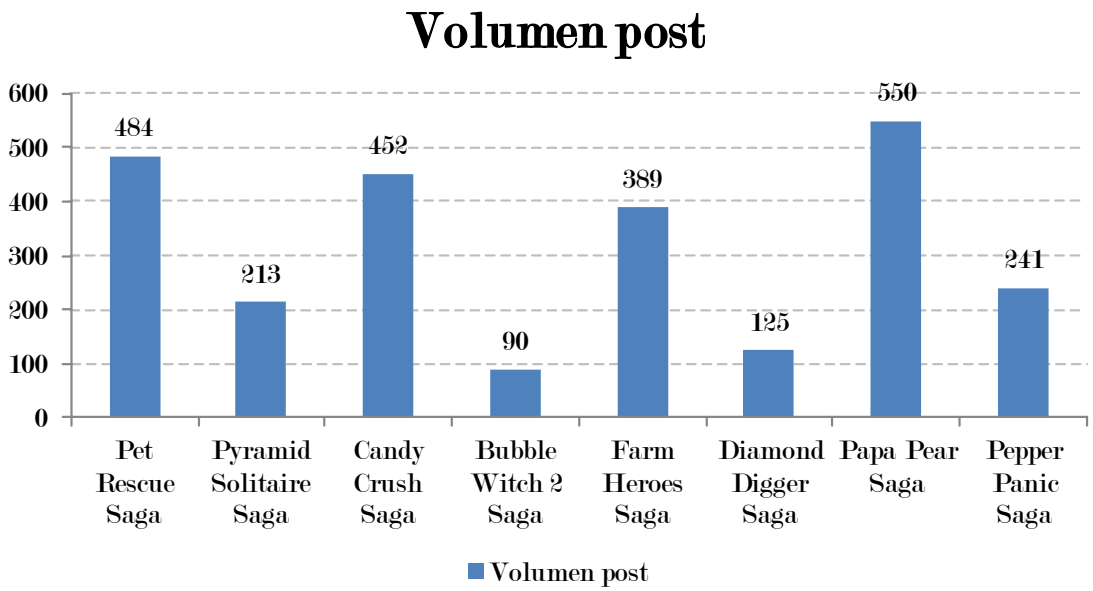
V.2. Descripción del escenario analizado

Para el caso práctico se van a analizar un conjunto de post de cada uno de los 8 juegos y las respuestas de los usuarios a esos post, proporcionados por el Institut Mines-Telecom SudParis con el que colabora mi tutor.

Utilizando la herramienta desarrollada en el proyecto, que tiene el detector del lenguaje integrado seleccionado en el capítulo anterior, se va a determinar el idioma de todos los post y todos los comentarios de los usuarios relativos a los 8 juegos. El objetivo es determinar el patrón de idioma que tienen los usuarios en función de los comentarios publicados y obtener toda la información disponible que sirva para mejorar las campañas de marketing digital o para promocionar de manera más eficiente algún producto.

V.2.1 Post analizados

Se han recopilado un total de de 2.544 post publicados por la empresa King Digital Entertaiment en las redes sociales de los 8 juegos. La distribución de los post según el juego es la siguiente:

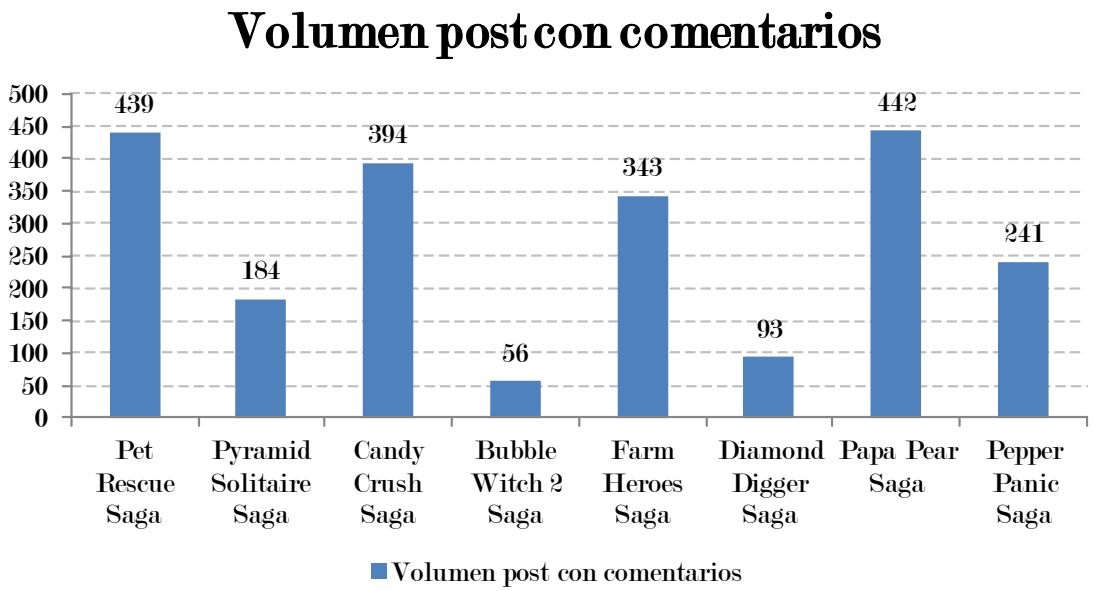


Gráfica 24: Volumen de post publicados por la empresa en las redes sociales

Se observa que la distribución de los post que se disponen no es uniforme por juego, se analizarán todos los comentarios relativos a cada juego en los siguientes apartados.

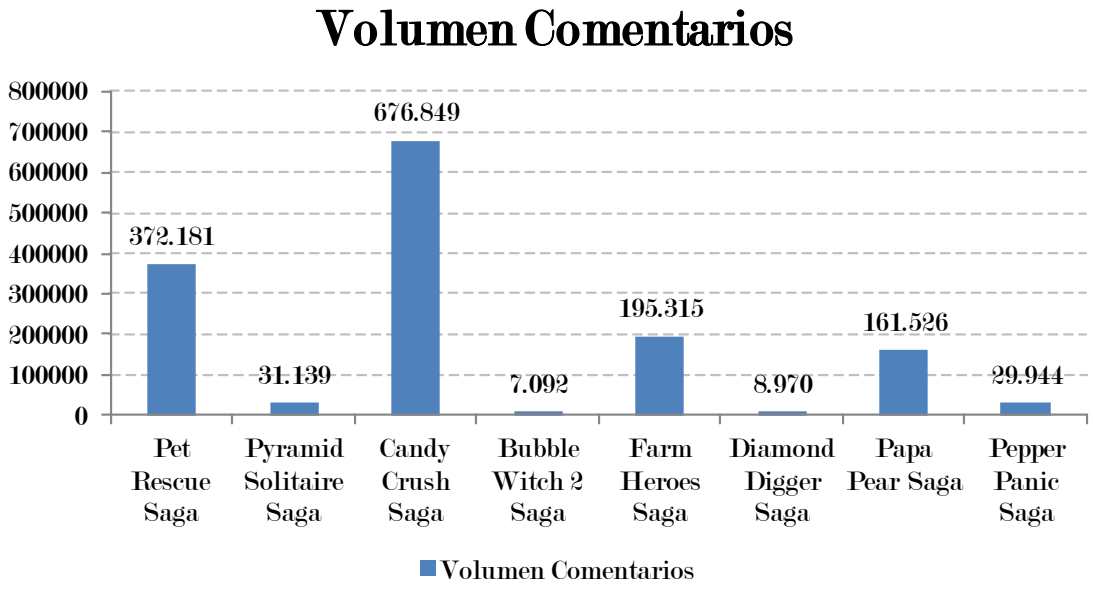
V.2.2 Respuestas usuarios analizadas

De los 2.544 post publicados por la empresa que se están analizando, 2.192 se han comentado por los usuarios en las redes sociales de todos los juegos. La distribución de los post comentados según el juego es la siguiente:



Gráfica 25: Volumen de post publicados por la empresa en las redes sociales con comentarios de usuarios

De todos los post comentados por los usuarios, se han generado 1.483.016 comentarios, en la gráfica siguiente se muestra la distribución de estos comentarios según el juego.



Gráfica 26: Volumen de comentarios de los usuarios según el juego

Se analizará el idioma de cada uno de los post y cada uno de los comentarios, de las gráficas anteriores se deduce que los comentarios de los usuarios no son proporcionales al volumen de post publicados, por tanto se analizarán en global y particularizando para cada juego

V.2.3 Test realizados

Se analizarán todos los idiomas de los post y de los comentarios de los usuarios, estudiando cada una de las distribuciones y las particularidades de cada juego.

Adicionalmente se presentan los resultados de los dos casos de usos reales que se han analizado.

- En el primero se analizará la distribución de los idiomas cuando el post publicado es en un idioma concreto.
- En el segundo se analizará la distribución de los idiomas de los comentarios de los usuarios, cuando los post que publica la empresa con una imagen o un video también tienen un texto, y se comparará con la distribución de los idiomas de los comentarios de los usuarios cuando los post que publica la empresa sólo tienen una imagen o un video.

V.3 Resultados del análisis

En este apartado se detallan todos los test realizados y los resultados obtenidos. Cada uno de ellos es clave para determinar el comportamiento de los usuarios en las redes sociales.

V.3.1 Distribución por post

Como se ve en la Gráfica 24 se dispone de 2.544 post. Los post se definen como una entrada, mensaje o publicación en una red social que puede consistir en un texto, opinión, comentario, enlace o archivo compartido.

Los post en este análisis los podemos clasificar en dos grupos:

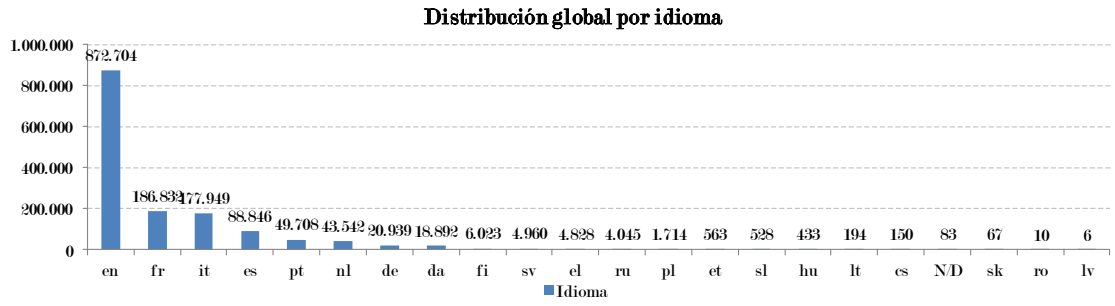
- Por el idioma de publicación. Al ser una empresa que tiene los usuarios de sus juegos repartidos por todo el mundo el idioma predominante en todas sus publicaciones es el Inglés. Concretamente para los post del estudio el idioma utilizado es el Inglés, excepto en dos post. Un post de Candy Crush se publica en Holandés y un post de Pyramid Solitaire se publica en Francés.

- Por el tipo de publicación. Aquellos que tienen una imagen o un video, en este tipo adicionalmente podrán llevar un texto complementando a la imagen o no.
  Estos dos últimos tipos de post se analizarán en los casos particulares (apartado V.3.4).

V.3.2 Distribución por comentarios

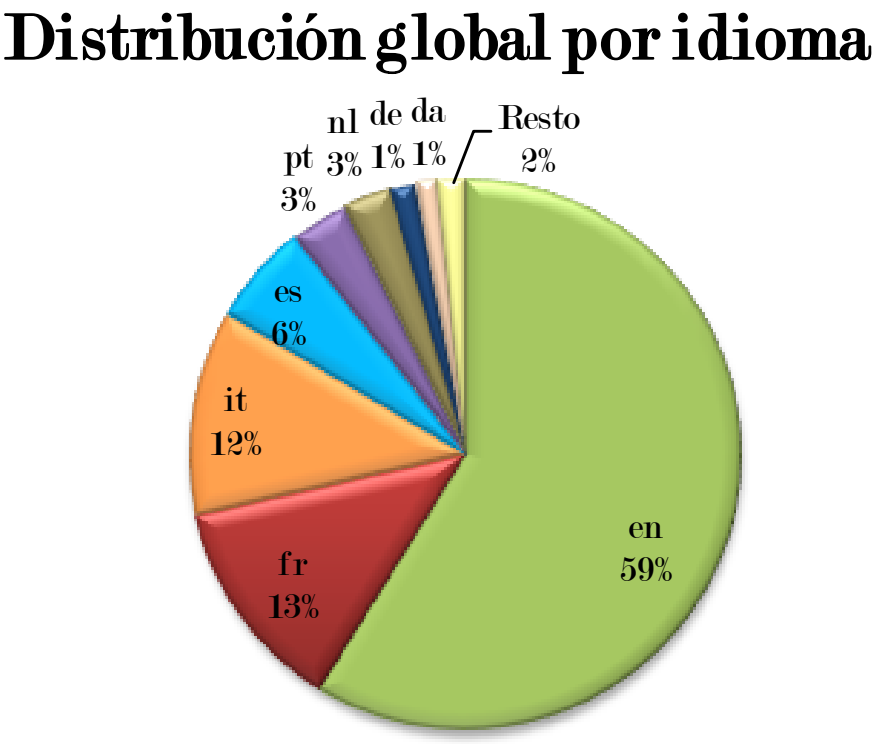
Se presenta la distribución por idioma de todos los comentarios, se han agregado por idioma todos los comentarios realizados por los usuarios.
Los idiomas representados en la gráfica son cs (checo), da (danés), de (alemán), el (griego), en (inglés), es (español), et (estonio), fi (finlandés), fr (francés), hu (húngaro), it (italiano), lt (lituano), lv (letón), nl (holandés), pl (polaco), pt (portugués), ro (rumano), sk (eslovaco), sl (esloveno) y sv (sueco).

Existe una gran dispersión de idiomas, aunque los idiomas más utilizados en los comentarios son el Inglés, el Francés y el Italiano.



Gráfica 27: Distribución de los comentarios por idioma

En la siguiente gráfica se expresa cual es el idioma en % de los comentarios realizados por los usuarios. Se han agregado en Resto todos los idiomas que representan menos del 1%. Aunque se han detectado 20 idiomas diferentes, en 8 idiomas se concentran el 98% de los comentarios. De manera general se observa que el inglés es el idioma predominante, por tanto se quiere acceder al mayor % población se debe publicar los post en Inglés.



Gráfica 28: Distribución por idioma de los comentarios de los usuarios

V.3.3 Distribución por juego

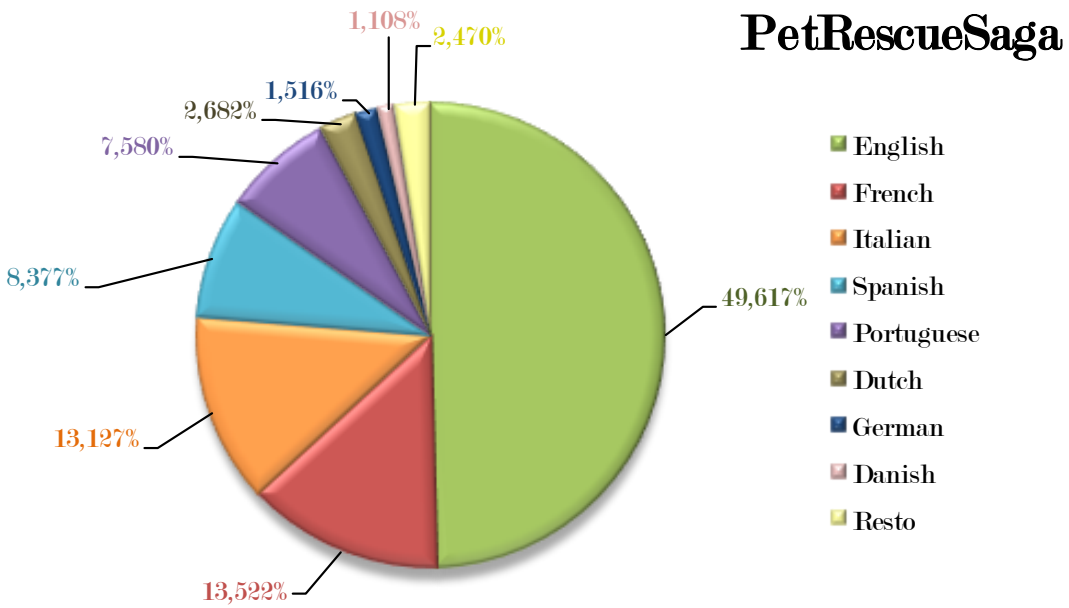
Se han separado los comentarios de cada juego y se han analizado los idiomas de cada juego por separado, para poder obtener conclusiones y ver si el comportamiento de los usuarios es distinto según el juego o se comportan todos de manera similar.

Las próximas 8 gráficas representan la distribución de los idiomas de todos los comentarios para cada uno de los juegos.

PetRescueSaga

Se han analizado 372.181 comentarios de los cuales el inglés supone el 50% de los comentarios, el Francés, el Italiano el Español y el Portugués están distribuidos equitativamente, suponen aproximadamente un 10% cada uno y el resto no llegan a un 10% entre todos.

Para el caso de este juego se puede ver que sudamerica en un mercado importante, Brasil y Argentina tienen un peso relevante de ahí que el Español y el Portugués tengan un peso tan importante y mayor que en la distribución global, acciones específicas para estos países o beneficios como vidas extras en el juego si tus vecinos se lo descargan, te permite publicitar el juego y que se hable en la calle de ello.

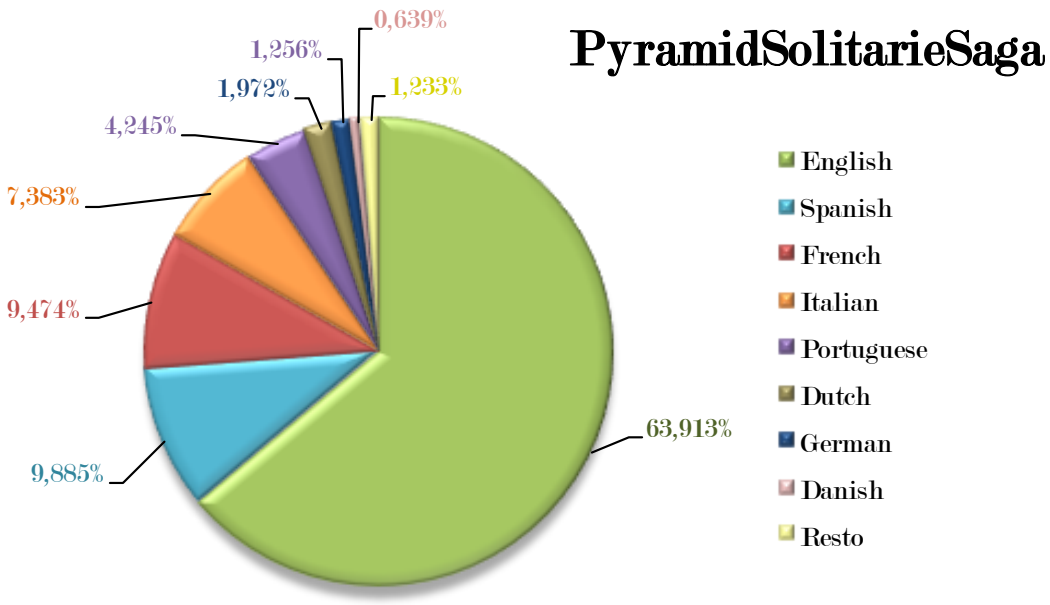


Gráfica 29: Distribución de los idiomas comentados por los usuarios en los post de PetRescueSaga

PyramidSolitaireSaga

Se han analizado 31.139 comentarios de los usuarios, el inglés supone el 64% de los comentarios, en este caso el segundo idioma más utilizado es el Español 9,8% a diferencia de la distribución global que es el Francés, el resto de idiomas superan levente el 25% de comentarios.

En este caso que los comentarios en Español y en Inglés superan el 73%, te permite lanzar encuestas en sólo dos idiomas y obtener el feedback de un volumen muy elevado, optimizando los costes de las encuestas.

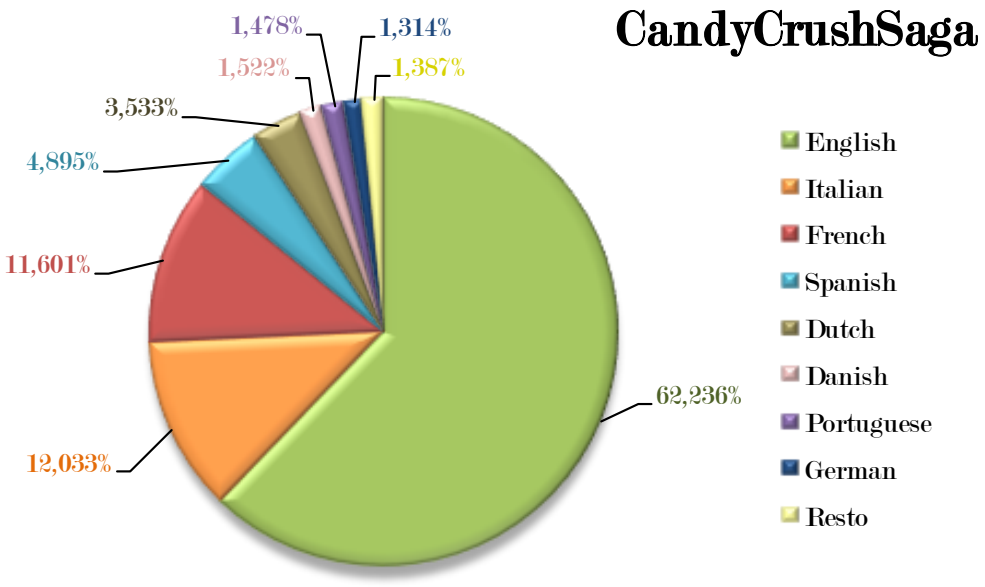


Gráfica 30: Distribución de los idiomas comentados por los usuarios en los post de PyramidSolitaireSaga

CandyCrushSaga

El juego más popular de la compañía y el más voluminoso en comentarios, se han analizado 676.849, de igual manera que el resto el inglés es el idioma más utilizado llegando en este caso hasta el 62,24% de los comentarios, junto con el Italiano y el Francés suponen aproximadamente un 85%.

Los Italianos y los Franceses son usuarios activos en Facebook, ya que el volumen de comentarios de este juego es mucho mayor que el resto, personalidades importantes de estas nacionalidades, como pueden ser cantantes o actores/actrices, que llegan de manera masiva a las personas, pueden ser una buena manera de promocionar el juego y además de servir como campaña publicitaria para todo el mundo, llegar a estos países de manera más importante por tener la misma nacionalidad.

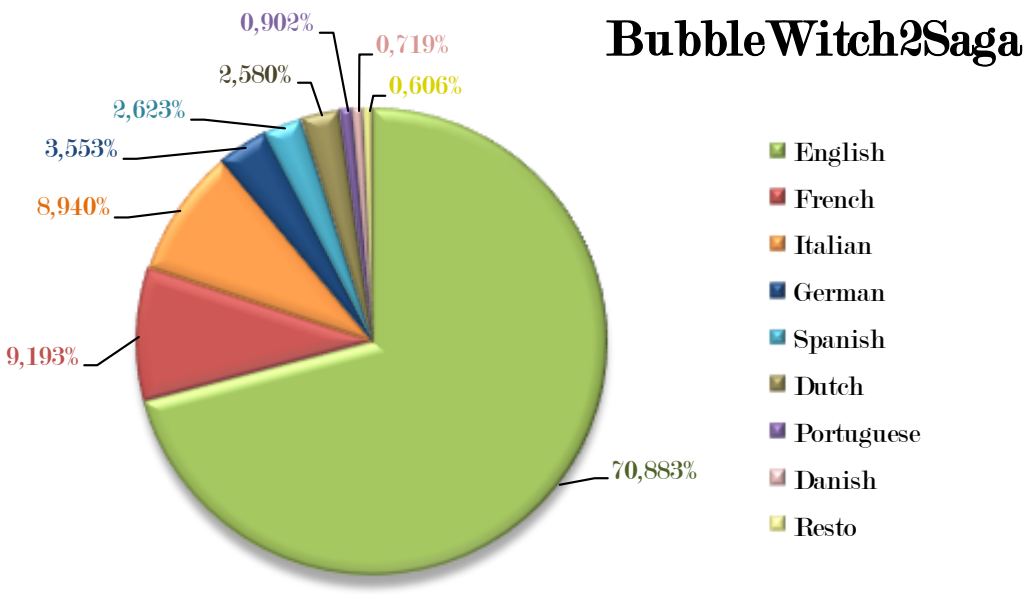


Gráfica 31: Distribución de los idiomas comentados por los usuarios en los post de CandyCrushSaga

BubbleWitch2Saga

En el juego se han analizado 7.092 comentarios, el inglés representa el % más alto de todos los juegos ascendiendo en este caso hasta el 70,88%. A diferencia de los otros juegos el alemán representa el 3,55%, se trata de un valor elevado comparado con el resto de juegos y con el valor global que el alemán suponía el 1%.

En este caso el Inglés es idioma clave para promocionar este juego, el volumen de comentarios en Inglés hace que no tenga sentido utilizar otro idioma, post o promociones en Inglés van a tener más efectividad que lanzar algo particular para un país.

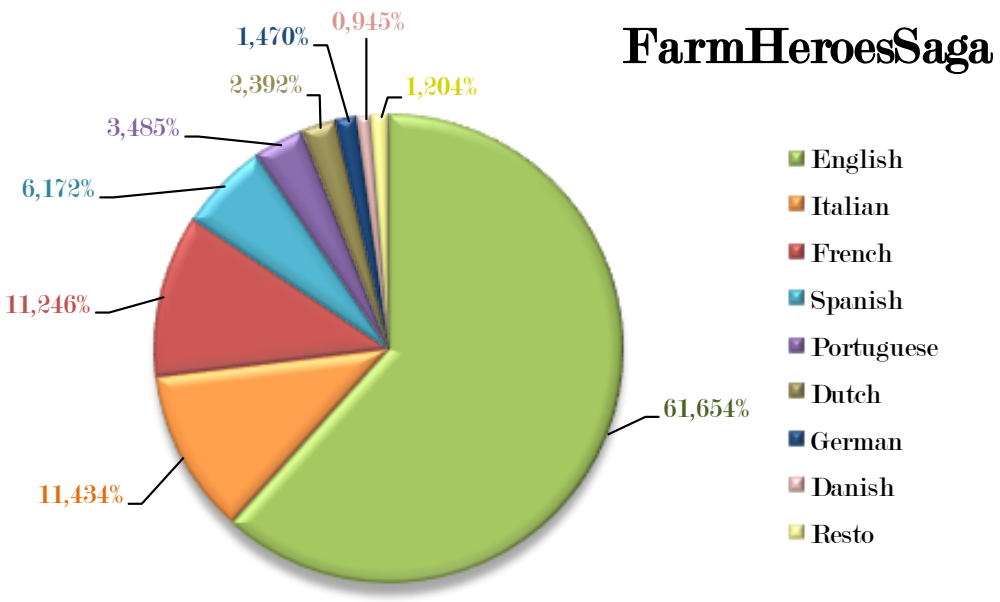


Gráfica 32: Distribución de los idiomas comentados por los usuarios en los post de BubbleWitch2Saga

FarmHeroeSaga

En el juego se han analizado 195.315 comentarios, en este caso el inglés supone el 61,65% y el italiano supera levente al Francés.

El volumen de comentarios en este caso es relevante, es el tercero después CandyCrushSaga y PetRescueSaga, en este caso es interesante Italia y Francia para lanzar una campaña publicitaria sobre una nueva evolución del juego o una promoción antes de lanzarla masivamente, esto permite obtener un feedback previo representativo. Gracias a la herramienta, te permite saber que para este juego utilizar Italia puede ser una buena opción.

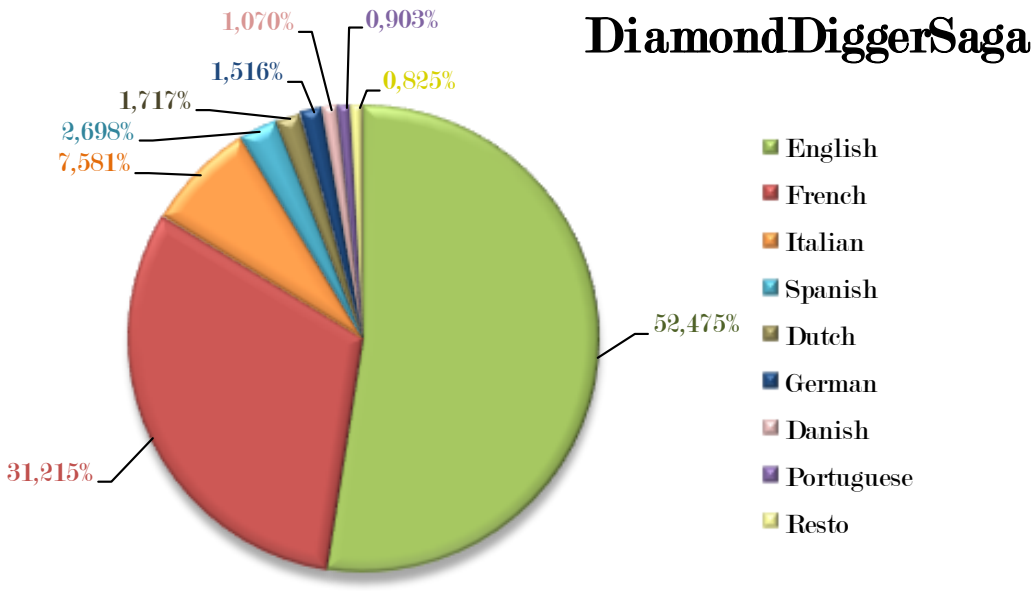


Gráfica 33: Distribución de los idiomas comentados por los usuarios en los post de FarmHeroeSaga

DiamondDiggerSaga

En el juego se han analizado 8.970 comentarios, de los cuales el inglés supone el 52,48% y es relevante que el Francés asciende hasta el 31,22% frente al 13% que supone en la gráfica agrupada de todos los comentarios. Esto es una particularidad de este juego, ya que un porcentaje tan elevado de un idioma distinto al Inglés no ocurre en ningún juego y es algo diferencial que se puede utilizar de manera positiva.

Para este juego debido al volumen de comentarios en Francés se deduce que los usuarios Franceses hacen un uso elevado del juego, tener páginas webs específicas en este idioma así como las versiones del juego en este idioma puede ser una herramienta interesante para fidelizar a este colectivo.

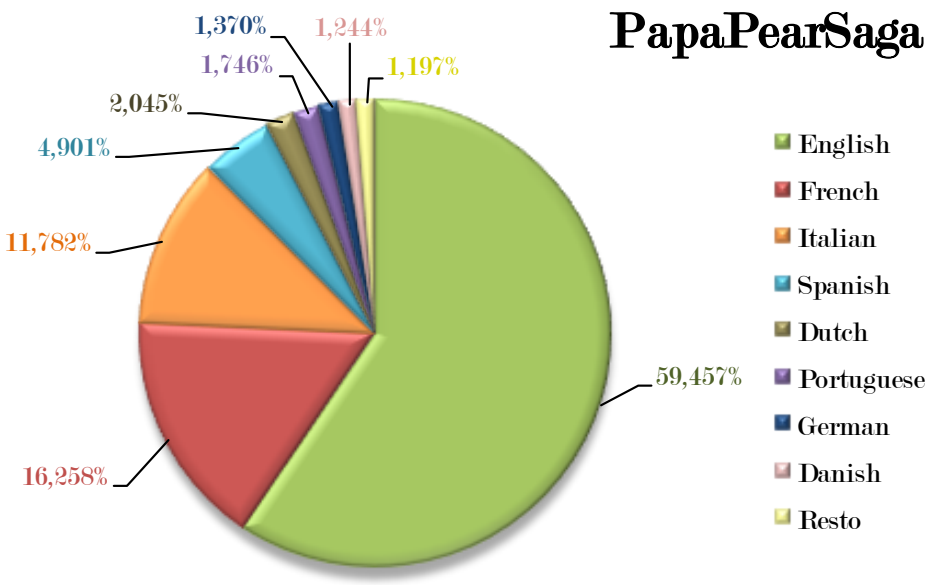


Gráfica 34: Distribución de los idiomas comentados por los usuarios en los post de DiamondDiggerSaga

PapaPearSaga

En el juego se han analizado 161.526 comentarios, de los cuales el inglés supone el 59,46% y junto con el Francés, el Italiano y el Español superan el 90% de los comentarios.

Para el caso de este juego los idiomas más populares después del Inglés son los Europeos, concretamente países mediterráneos, son países que reciben gran cantidad de turistas en verano por lo que campañas publicitarias en esa época puede suponer mayor promoción para el juego.

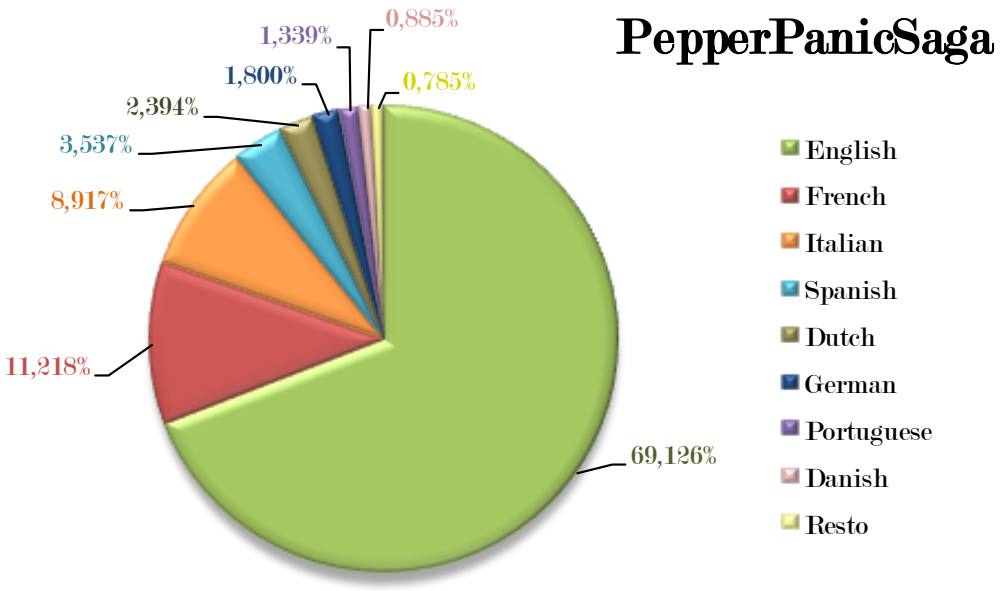


Gráfica 35: Distribución de los idiomas comentados por los usuarios en los post de PapaPearSaga

PepperPanicSaga

En el juego se han analizado 29.944 comentarios, de los cuales el Inglés supone el 69,13%, el Francés representa el 11,22%, el Italiano el 9,912% y el Español el 3,54%.

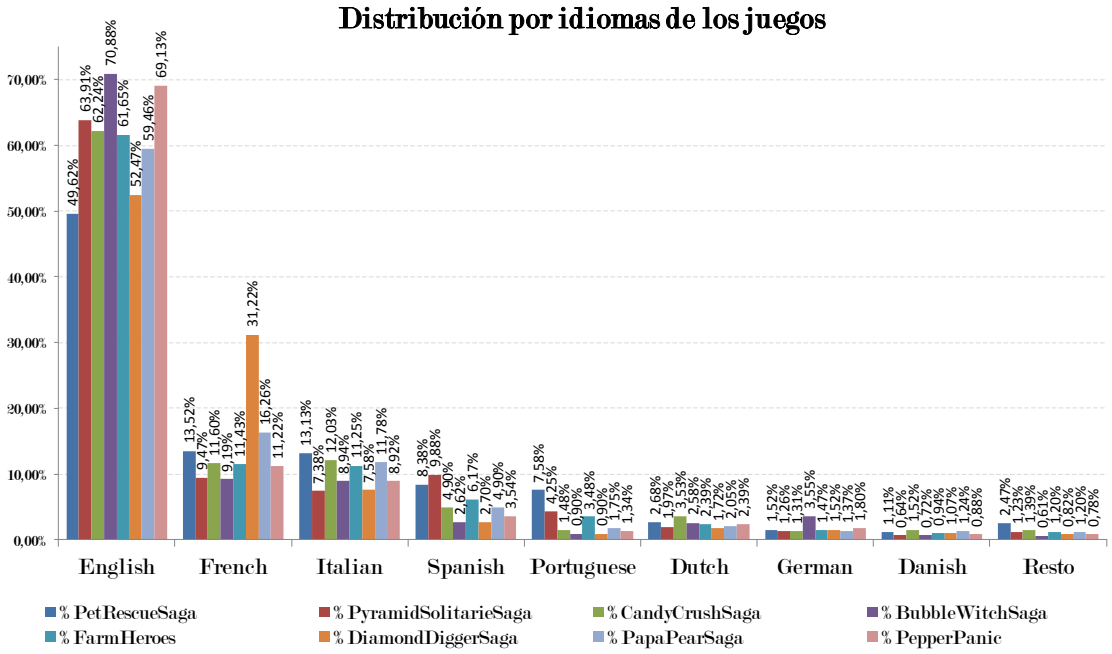
En este caso el Inglés tiene un peso muy importante y si no se pueden abarcar todos los idiomas, post en Inglés te van a garantizar que les llega a un porcentaje muy amplio.



Gráfica 36: Distribución de los idiomas comentados por los usuarios en los post de PepperPanicSaga

Como resumen en la siguiente gráfica se muestra una comparativa de todos los juegos y todos los idiomas, de manera general el idioma predominante en todos los juegos es el Inglés, superando de manera considerable, en más del doble, al resto. De este análisis se observa que el público prefiere comunicarse en Inglés y es el idioma más popular entre los usuarios por tanto si se pretende abarcar a un público amplio el Inglés es el idioma que se debe utilizar para esta empresa.

El comportamiento es similar en todos los juegos, el inglés se encuentra entre el 49,61% en PetRescueSaga y 70,88% en BubbleWitch2Saga, sin embargo el resto de idiomas permite diferenciar y particularizar para cada juego. Estas particularidades, según el idioma en el que se comentan los post, puede servir para ser más efectivo en un juego en concreto o minimizar los costes siendo más efectivos. Esto en el entorno social que nos encontramos es clave para cualquier empresa y supone poco esfuerzo frente a los beneficios que presenta.



Gráfica 37: Distribución en % de los idiomas comentados por los usuarios según el juego

V.3.4 Casos Particulares

En este apartado se pretende mostrar los casos particulares en los que se puede utilizar la herramienta y la utilidad de la misma.

V.3.4.1 Post en idioma diferente

Se hace zoom en el idioma de las respuestas que dan los usuarios a post en un idioma distinto al inglés. En el apartado anterior se han analizado las respuestas en global obteniendo unas conclusiones, sin embargo, si sólo nos centramos en las respuestas de los post que no están en inglés las conclusiones varían. A continuación se detalla el análisis.

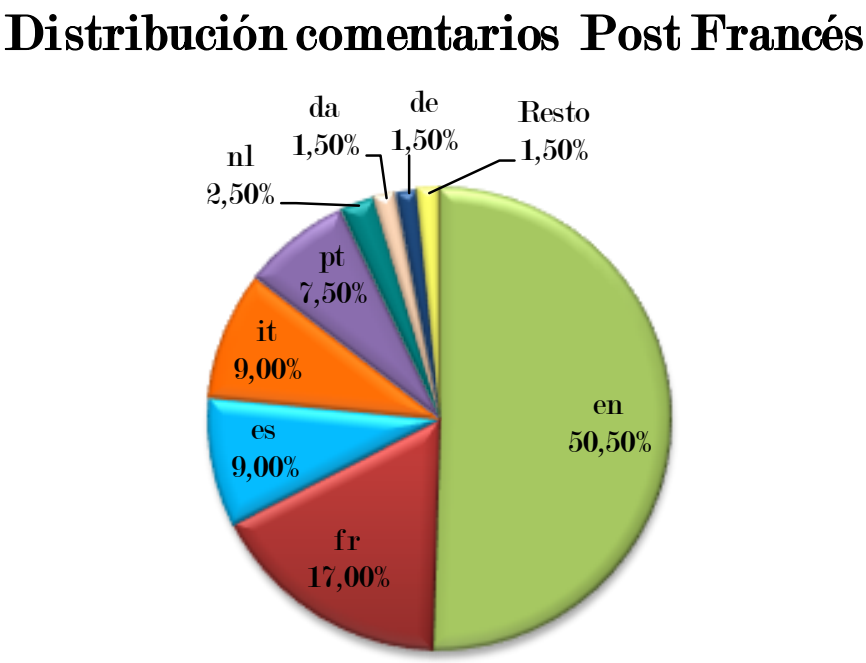
Post Francés

En la red social de Facebook del juego de Pyramid Solitaire Saga, se ha publicado el siguiente post en Francés:

Figura 6: Post en Francés en Facebook

Este post ha generado 200 respuestas diferentes, de las cuales el 50,5% han sido en Inglés y el 17% en Francés. Si comparamos estos datos con la distribución del juego con todos los comentarios detallado en el apartado V3.3, se aprecia que el porcentaje de Inglés disminuye pasando de 63,91% cuando los post son en Ingles

frente al 50,50% cuando el post es otro idioma, concretamente en Francés. Con el idioma Francés ocurre lo contrario se produce un incremento en este caso pasando del 9,47% en la distribución del juego con todos los comentarios a un 17%, se ha producido un incremento de casi el doble en este idioma, manteniéndose en los mismos porcentajes el resto de idiomas.



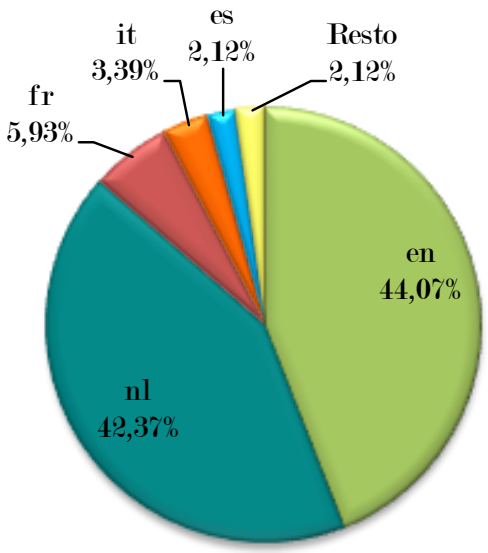
Gráfica 38: Distribución por idiomas de los comentarios de los usuarios a un post en Francés

Post Holandés

En la red social de Facebook del juego de Candy Crush Saga, se ha publicado el siguiente post en Holandés.

Figura 7: Post en Holandés en Facebook

Este post ha generado 236 respuestas diferentes, de las cuales el 44,07% han sido en Inglés y el 42,37% en Holandés. Se ha cambiado drásticamente la distribución que se mostraba en el apartado V3.3 del juego Candy Crush pasando a ser el Holandés un idioma con un porcentaje muy elevado del 3,55% inicial a un 42,37% en este caso concreto, el resto de idiomas se reduce al 17%.



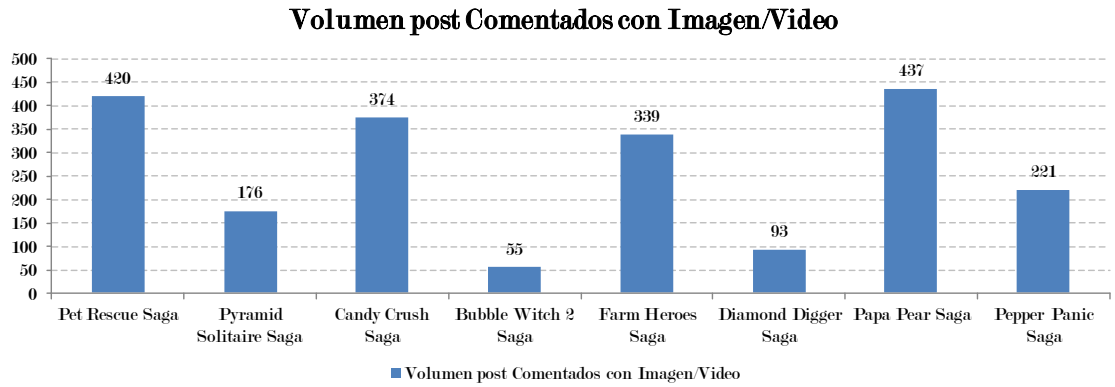
Gráfica 39: Distribución por idiomas de los comentarios de los usuarios a un post en Holandés

De este caso se observa que el idioma en el que se publican los post influye de manera cuantitativa en las respuestas de los usuarios. Si se requiere promocionar algún producto en un lugar geográfico concreto publicar en el mismo idioma va a suponer un aumento de las respuestas en ese idioma, en el caso de uso se ha incrementado el Francés y el Holandés.

V.3.4.2 Influencia de incluir un texto en post con video/imagen

En este caso se hace zoom en la respuesta que dan los usuarios a un post que sólo tiene un video/imagen frente a las respuestas que dan cuando el post además del video/imagen tiene un texto.

De todos los post que se han analizado en el apartado V.3.3, tenemos 2.115 post que contienen una imagen o un video, la distribución de los mismos según el juego está representada en la siguiente gráfica.

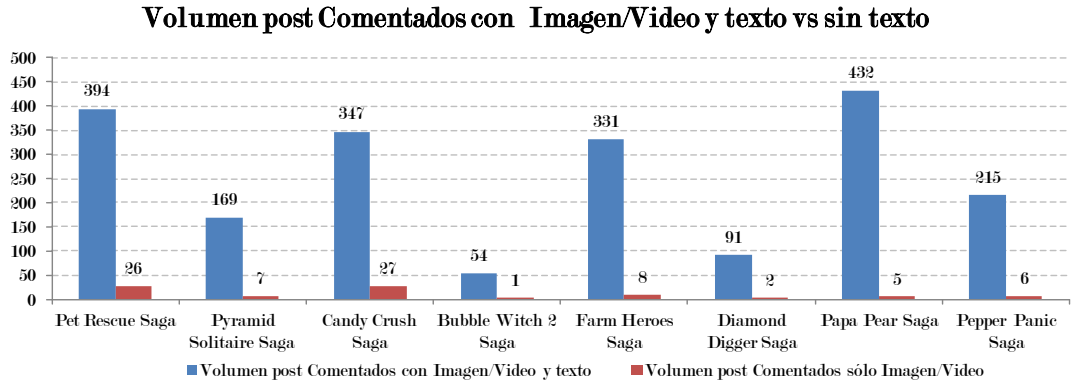


Gráfica 40: Volumen de post con imagen o video

De los 2.115 post que se van a analizar en este apartado tenemos:

- 2033 la empresa ha publicado la imagen/video y han incluido un texto
- 82 post se han publicados sólo con la imagen/video.

El número de post analizados de cada juego se puede observar en la siguiente gráfica.



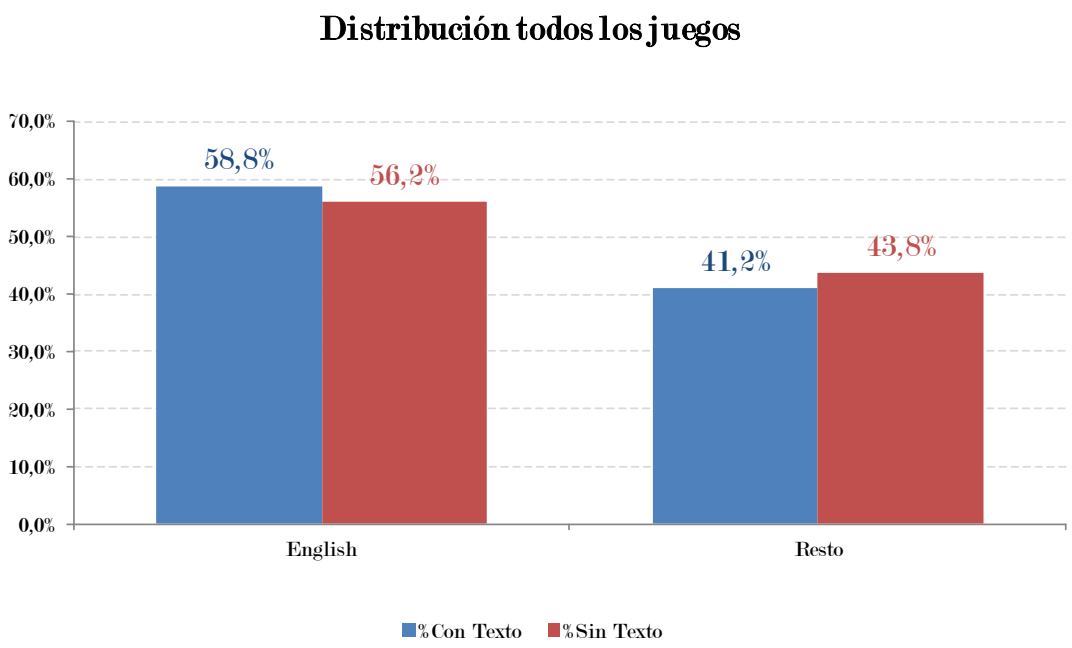
Gráfica 41: Volumen de post con imagen/video y texto vs post sólo con imagen/video

Estos post que se han mostrado en la distribución anterior, ha supuesto una repercusión Facebook con la siguiente repercusión:

- Los 2.033 post que la empresa ha publicado con imagen/video y un texto han generado 1.401.790 comentarios de los usuarios
- Los 82 post que se han publicados sólo con la imagen/video han generado 58.928 comentarios de los usuarios.

Se ha analizado con la herramienta el idioma de estos comentarios, en la siguiente gráfica se muestran la distribución por idiomas. En las columnas azules se puede ver el porcentaje de cada idioma de los comentarios que se han generado de los post con imagen/video y texto y en las columnas rojas se muestra el porcentaje de cada idioma de los comentarios que se han generado de los post sólo con imagen/video.

Otro dato importante para el análisis es que el texto que la empresa ha incluido en los post es en todos los casos en Inglés, por tanto con este análisis se pretende mostrar si en un post con una imagen/video incluir un texto en Inglés condiciona a los usuarios a contestar en el mismo idioma o si se comporta de manera distinta si no se incluye nada.



Gráfica 42: Distribución por idioma de los comentarios de los usuarios a los post con imagen/video y texto vs sin él.

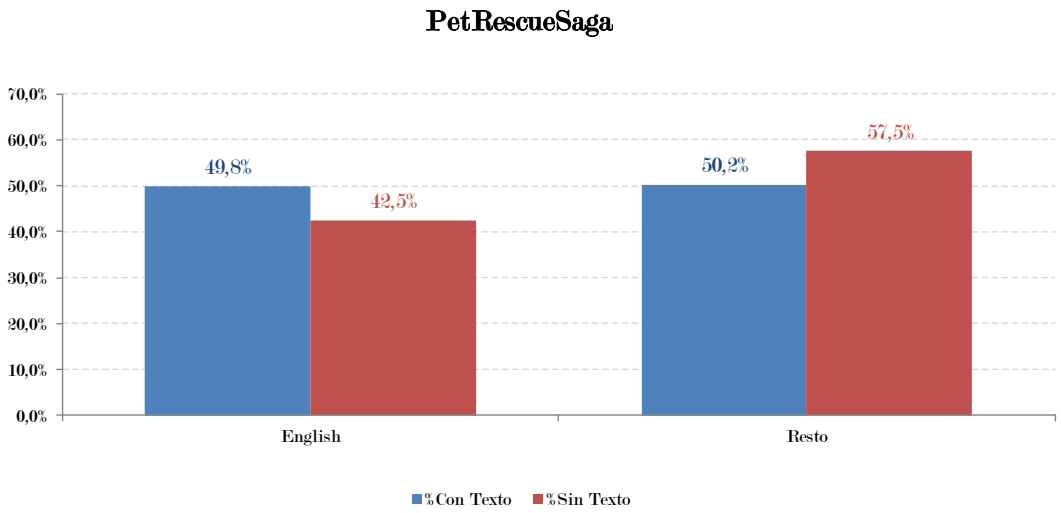
En aquellos post sin texto, en los cuales el usuario no tiene ninguna referencia el porcentaje del resto de idiomas incrementa casi en 3 pp, esto quiere decir que si queremos fomentar un debate a través de una imagen en un idioma en concreto, se recomienda incluir alguna frase en el idioma que se desea.

Se ha realizado el análisis desglosado por cada uno de los juegos y en las siguientes gráficas se muestran los resultados.

PetRescueSaga

En el juego se han analizado 351.458 comentarios con texto adicional, de los cuales el 49,8% es Inglés y 13.443 sin texto, de los cuales el 42,5% es en Inglés. El resto de idiomas el porcentaje comentarios es mayor cuando los post no contienen textos adicionales.

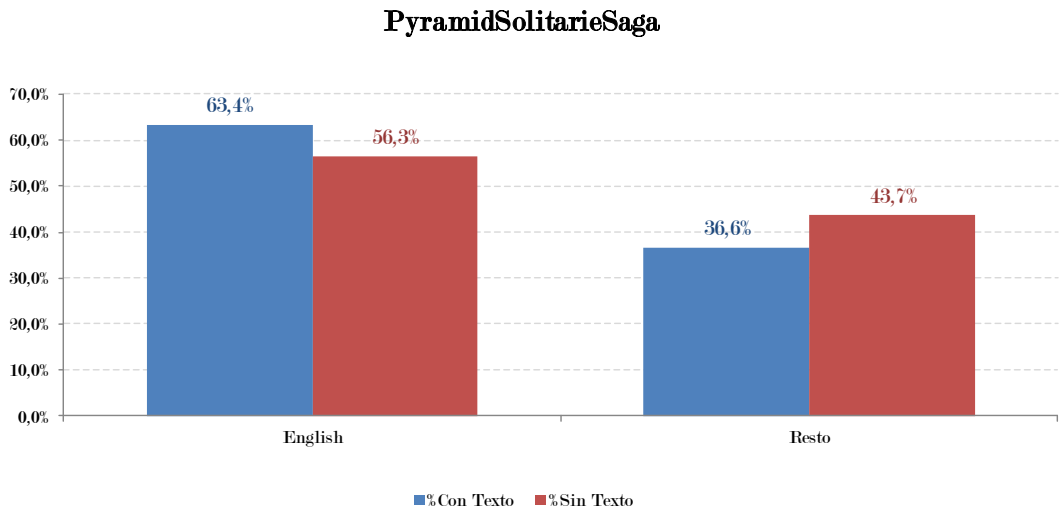
Este caso es particular, ya que a diferencia de la distribución global el resto de idiomas cuando no se incluye ningún texto supera notablemente al Inglés. En este juego si sería interesante no incluir ningún texto y que se genere un debate de la imagen en múltiples idiomas, esto proporciona mayor publicidad y promoción para lo que se quiera mostrar, ya que el público que puedes abarcar es mayor si la gente deja comentarios en múltiples idiomas.



Gráfica 43: Distribución por idiomas de los comentarios de los usuarios a los post del juego PetRescueSaga con imagen/video y texto adicional vs sin él

PyramidSolitaireSaga

En el juego se han analizado 27.764 comentarios con texto adicional, de los cuales el 63,4% es Inglés y 1.614 sin mensaje, de los cuales el 56,3% es en Inglés. El resto de idiomas el porcentaje comentarios es mayor cuando los post no contienen textos adicionales.

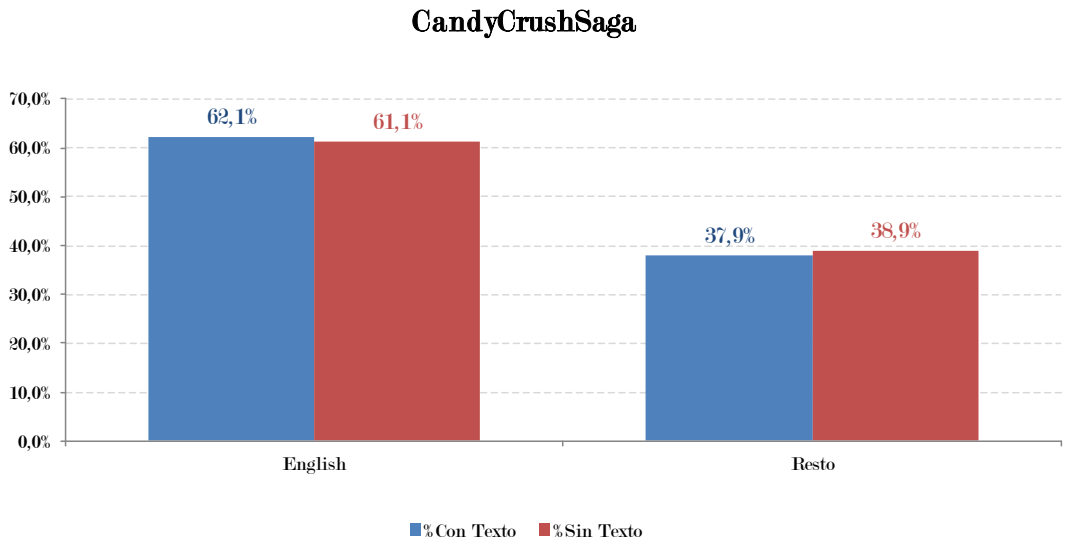


Gráfica 44: Distribución por idiomas de los comentarios de los usuarios a los post del juego PyramidSolitaireSaga con imagen/video y texto adicional vs sin él

CandyCrushSaga

En el juego se han analizado 630.589 comentarios con texto adicional, de los cuales el 62,1% es Inglés y 34.225 sin texto, de los cuales el 61,1% es en Inglés.

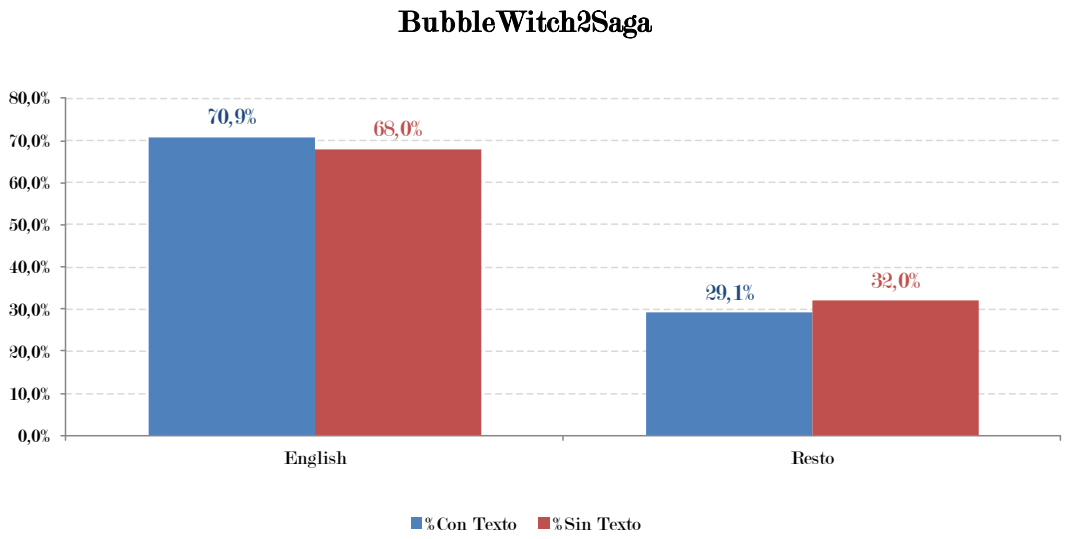
Para este juego con los resultados obtenidos, no sería efectivo incluir un texto adicional para fomentar el uso de un idioma, ya que la diferencia es mínima, indistintamente se utiliza el Inglés por los usuarios. Para incrementar el uso de otros idiomas se debe buscar otras alternativas, como por ejemplo, introducir un texto en un idioma distinto al Inglés como se ha visto en el apartado anterior.



Gráfica 45: Distribución por idiomas de los comentarios de los usuarios a los post del juego CandyCrushSaga con imagen/video y texto adicional vs sin él

BubbleWitch2Saga

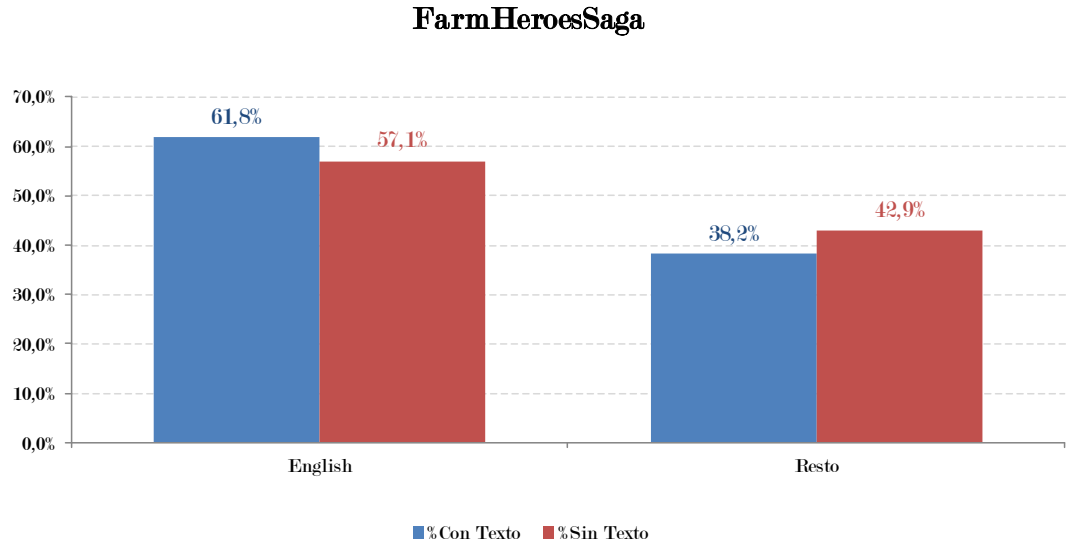
En el juego se han analizado 6.925 comentarios con texto adicional, de los cuales el 70,9% es Inglés y 25 sin texto, de los cuales el 68% es en Inglés. Aunque no tenemos mucha muestra se aprecia la misma tendencia que en la distribución global.



Gráfica 46: Distribución por idiomas de los comentarios de los usuarios a los post del juego BubbleWitch2Saga con imagen/video y texto adicional vs sin él

FarmHeroeSaga

En el juego se han analizado 187.584 comentarios con mensaje adicional, de los cuales el 61,8% es Inglés y 7.140 sin mensaje, de los cuales el 57,1% es en Inglés. En este caso la mayoría de los idiomas incrementan su porcentaje cuando no tienen comentario adicional, dentro del resto de idiomas en este caso el Español destaca teniendo el mayor incremento pasando de 6% a 9,8%.



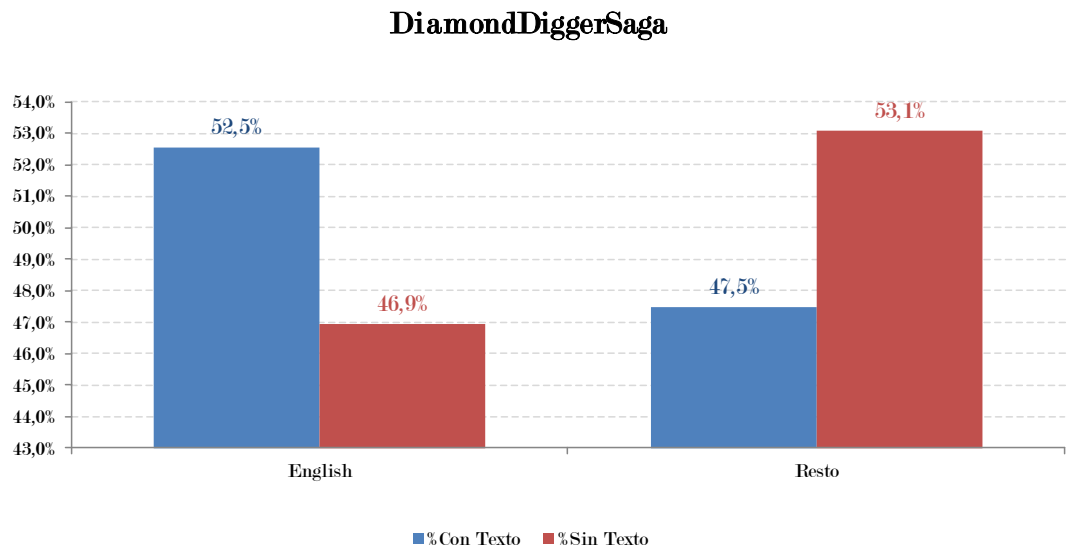
Gráfica 47: Distribución por idiomas de los comentarios de los usuarios a los post del juego FarmHeroeSaga con imagen/video y texto adicional vs sin él

DiamondDiggerSaga

En el juego se han analizado 8.872 comentarios con texto adicional, de los cuales el 52,5% es Inglés y 98 sin texto, de los cuales el 46,9% es en Inglés.

En este caso en particular se aprecia que cuando no se incluye un texto el Inglés no es el idioma principal, esto quiere decir que los usuarios de Facebook que comentan este juego prefieren hacerlo en otro idioma.

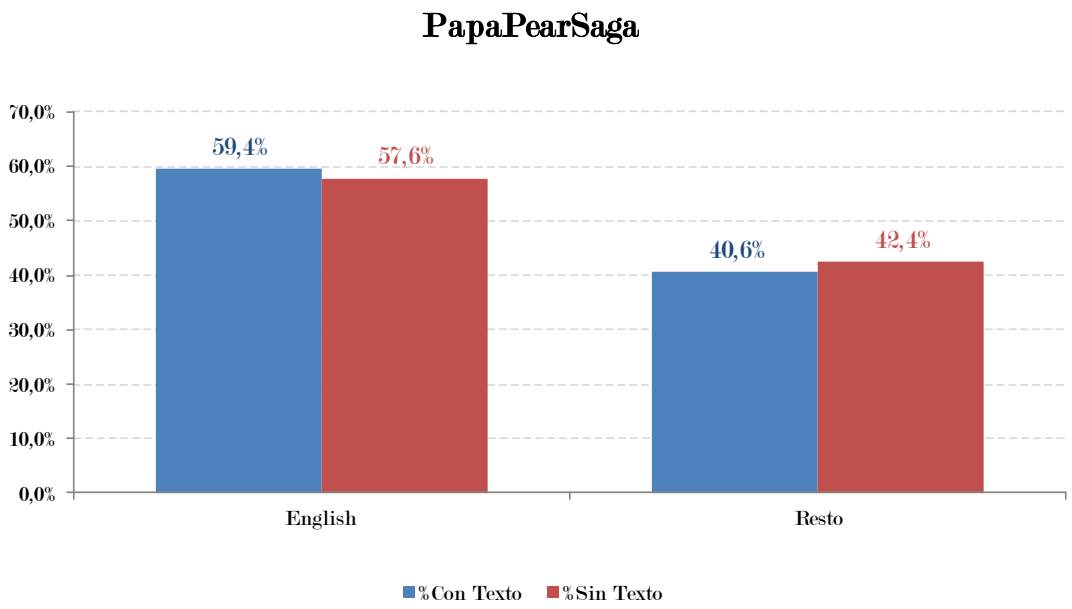
Este caso es un claro ejemplo para realizar campañas de Facebook con sólo imágenes, esto no condiciona a la gente a responder en el mismo idioma y puede ser más accesible a todos los usuarios ya que cada uno comenta lo que le sugiere en su idioma, llegando así a un mayor número de usuarios.



Gráfica 48: Distribución por idiomas de los comentarios de los usuarios a los post del juego DiamondDiggerSaga con imagen/video y texto adicional vs sin él

PapaPearSaga

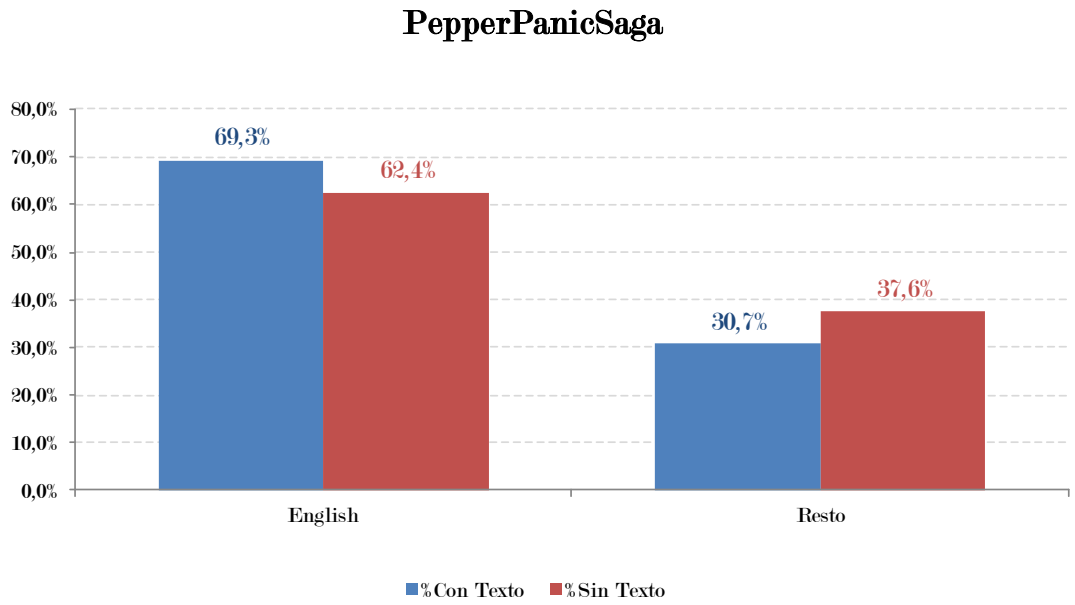
En el juego se han analizado 159.532 comentarios con texto adicional, de los cuales el 59,4% es Inglés y 1.505 sin texto, de los cuales el 57,6% es en Inglés. En este caso la diferencia de ambos casos es muy poca si se tiene que priorizar algún juego éste no sería el método para este juego.



Gráfica 49: Distribución por idiomas de los comentarios de los usuarios a los post del juego PapaPearSaga con imagen/video y texto adicional vs sin él

PepperPanicSaga

En el juego se han analizado 29.066 comentarios con texto, de los cuales el 69,3% es Inglés y 878 sin texto, de los cuales el 62,4% es en Inglés. El resto de idiomas todos superan en porcentaje cuando los post no tienen texto adicional en Inglés.



Gráfica 50: por idiomas de los comentarios de los usuarios a los post del juego PepperPanicSaga con imagen/video y texto adicional vs sin él

En la mayoría de los casos cuando no se publica un post con un texto no se condiciona a los usuarios a responder en el mismo idioma, y por tanto se observa que el porcentaje de comentarios en idiomas distintos al inglés se incrementa notablemente.

Esto depende del efecto que se quiera producir se puede añadir un texto en un idioma o dejarlo abierto. En el caso de querer una respuesta interactiva de todos los usuarios, si se produce en el mismo idioma te permite que todos puedan comentar, en este caso añadir un comentario propicia el comentar en ese mismo idioma. Aunque el Inglés es un idioma muy extendido con este tipo de post habrá usuarios a los que no se podrá involucrar.

En el otro caso de querer dejarlo abierto, cada usuario va comentar la imagen o el video según lo sienta, si se pretende llegar al máximo de usuarios cada uno va a comentarlo en el idioma que le sea más sencillo expresarse ya que en el post no condiciona con ningún comentario, la accesibilidad a los usuarios es total, no existe ninguna barrera lingüística que evite comentar lo que están viendo los usuarios.

Por tanto si se quiere publicitar algo que se quiera hacer a nivel global, no incluir ningún texto en la imagen generará un foro más rico en múltiples idiomas que será accesible y compartido por un volumen mayor de usuarios.

V.4. Conclusiones

En este apartado se pretende resumir las principales conclusiones obtenidas de los análisis realizados con la herramienta de detección de idiomas en Facebook.

Se han realizado tres análisis distintos para ver el comportamiento de los usuarios:

- Un análisis global de todos los comentarios proporcionados de Facebook de los 8 juegos principales de King.
- Un análisis particular cuando se publica un post en un idioma distinto al Inglés.
- Un análisis particular cuando se publica un post con una imagen/video con un texto adicional o sin él.

Del **primer análisis** se ha obtenido la distribución por idiomas para cada juego, el idioma predominante es el Inglés en todos los casos, por tanto, si el objetivo es abarcar a un público amplio fomentar el uso del Inglés va a ser clave para esta empresa.

De este análisis las conclusiones más interesantes las obtenemos de los segundos idiomas, ya que eso varía dependiendo del juego de manera sustancial, como se ha analizado caso a caso en el apartado correspondiente. Por tanto, la herramienta se ha utilizado para ver esas particularidades y poder tener propuestas personalizadas para cada juego, minimizando costes en campaña de publicidad, permitiendo testar a un público representativo previo a un lanzamiento.

El **segundo análisis** se centra en los post con un idioma concreto y la conclusión principal es que el idioma en el que se publica un post impacta directamente en la respuesta de los usuarios, promoviendo el uso del mismo idioma en los comentarios de los usuarios.

Es clave, para captar la atención de un colectivo o promover la participación utilizar el idioma que hable dicho colectivo, de esta manera vamos a facilitar la lectura del post, que tenga una mayor aceptación, con esto se fomentan los comentarios en el mismo idioma y que se involucre a todos los usuarios ya que no hay barreras lingüísticas que interfieran.

El **tercer análisis** se centra en los post que tienen un imagen/video y como varían las respuestas de los usuarios si en esos post se añade un texto. La conclusión principal es que la respuesta de los usuarios si tiene un texto se realiza en el mismo

idioma, sin embargo si no existe ningún texto responden en el idioma nativo. Por tanto, para acceder a más usuarios se observa que la mejor opción es no incluir ningún texto complementario, de esta manera cada usuario responderá en el idioma deseado y se accederá a un mayor número de personas, fomentando así la participación en el foro.

De manera global se puede ver que el idioma de los post en Facebook es un elemento clave para despertar la respuesta o colaboración de un grupo de usuarios concreto. Las conclusiones obtenidas corresponden a los casos de uso estudiados, son dos ejemplos de cómo se puede utilizar la herramienta según el objetivo que se quiera obtener.

Estas conclusiones se han obtenido para el caso práctico estudiado que son los post y comentarios de los juegos King, sin embargo la herramienta de detección de Idiomas en Facebook se puede aplicar a cualquier empresa que tenga una cuenta en Facebook con los post que publican y los comentarios que reciben de los usuarios. Esto se puede utilizar en distintos ámbitos:

- Análisis previo a campañas de marketing, esto permite analizar la distribución global de lo que se ha publicado y enfocar las campañas a un colectivo en un idioma, o utilizar el idioma mayoritario para llegar al mayor publico posible.
- Análisis post campañas de marketing, se puede analizar la efectividad de las mismas, el volumen de participación según el idioma, compararlo con lo estimado.

Se pretende con esta herramienta que las empresas tengan información adicional para optimizar costes y mejorar campañas de marketing, ya que no todos los usuarios de Facebook se comportan igual.

Con un conjunto de post y comentarios se puede obtener la distribución de los idiomas y analizar todas las particularidades de cada empresa, producto o servicio, para poder personalizar la campaña de marketing digital, el lanzamiento de un nuevo producto o una promoción.

La redes sociales y en particular Facebook es un método muy accesible a todo el mundo, con gran repercusión social y analizar las particularidades de los usuarios a cada producto puede ser la diferencia entre el éxito o el fracaso. El idioma es clave para comunicarse con los clientes y mientras más detalles de las respuestas de los usuarios más nos podremos acercar a los mismos.

CAPÍTULO VI: TRABAJOS FUTUROS

En este capítulo se pretende dar unas pinceladas de las posibles evoluciones que presenta la herramienta.

VI.1 Introducción

El objetivo del proyecto era crear un programa que automatice la herramienta seleccionada para identificar automáticamente el idioma de los comentarios de Facebook.

La herramienta que se ha desarrollado para este proyecto fin de carrera presenta distintas evoluciones que pueden ser desarrolladas a futuro para dar prestaciones complementarias.

Con la misma filosofía del proyecto de darle a las empresas información adicional de los usuarios que permita minimizar costes y ser más efectivos en los siguientes apartados se presentan posibles evoluciones.

VI.2 Ampliación Redes Sociales

En este caso nos hemos centrado en el análisis de Facebook, pero como se presentó en el primer capítulo existe un amplio abanico de redes sociales cada vez con más relevancia en la sociedad.

Una posible evolución es ampliar esta herramienta no sólo con los comentarios de Facebook sino a los comentarios de otras redes sociales, redes sociales como LinkedIn está pensada para un ámbito profesional que puede proporcionar a las empresas un análisis de los usuarios que comentan en su red social y el perfil y las conclusiones pueden ser completamente diferentes a las obtenidas, ya que el uso de ambas redes sociales es diferente.

Será necesario evaluar si el tipo de texto es el mismo o si por el contrario tiene características distintas.

Con esta evolución si las empresas tienen presencia en distintas redes sociales se pueden realizar comparativas de las distintas redes sociales para poder personalizar en cada caso.

VI.3 Interfaz Gráfico

Dotar a la herramienta de un interfaz gráfico que permita sacar las gráficas mostradas en el proyecto en un interfaz gráfico de manera sencilla para los usuarios.

Definiendo gráficas estándar con los datos introducidos, en las que se muestren distribuciones por idioma globales de post, comentarios… o gráficas particulares según el caso de uso que esté interesado el usuario de la herramienta.

CAPÍTULO VII: PRESUPUESTO ECONÓMICO

VII.1. Introducción

El objetivo es definir el conjunto de actividades y desarrollos necesarios para la realización de "Herramienta de Detección de Idiomas en Facebook".

En este capítulo se presenta la estimación de tiempos y recursos necesarios para la realización de las tareas descritas incluyendo la valoración económica correspondiente a dicha estimación.

VII.2. Descripción de la solución

Las actividades que comprenden el proyecto son análisis de las herramientas existentes y desarrollo de aplicación acorde al mejor detector, pruebas de la aplicación y entrega de documentación.

VII.2.1. Desarrollo aplicación

En el lenguaje de programación Java, se desarrollará una aplicación que permita identificar de manera automática y recurrente los mensajes que los usuarios dejan en las redes sociales.

Se evaluarán los distintos detectores del mercado y se integrará el mejor dentro del programa desarrollado, para dar una solución global.

VII.2.2. Pruebas e2e

Definición y ejecución del plan de pruebas:

- De autocertificación a ejecutar por el desarrollador, este es el conjunto de pruebas que realiza el desarrollador para verificar el correcto funcionamiento.
- End to end con el detector seleccionado, así como evaluación del rendimiento de la herramienta.

Definición y puesta en marcha del entorno necesario para la ejecución de los protocolos de pruebas. Esto supone disponer de PCs con JRE (Java Runtime Environment), antiguamente conocida como Java Virtual Machine (JVM) y un repositorio con los post que nos proporciona la universidad.

Reporte de los resultados del plan de pruebas y seguimiento de los bugs reportados hasta la resolución de los mismos.

Identificar y reproducir los bugs identificados en cada versión.

VII.2.3. Documentación

Documentación de todas las clases del programa con:

- Nombre de la clase, descripción general, número de versión, nombre de autores.
- Documentación de cada constructor o método (especialmente los públicos) incluyendo: nombre del constructor o método, tipo de retorno, nombres y tipos de parámetros si los hay, descripción general, descripción de parámetros (si los hay), descripción del valor que devuelve.

Archivo Readme, que incluirá título del proyecto, descripción, versión, cómo arrancar el proyecto, autores e instrucciones para los usuarios.

Versión final de los planes de prueba y resultados de las pruebas de autocertificación y end to end.

VII.2.4. Beneficios de la solución

Los beneficios más destacados de la solución propuesta son:

- Fácil de usar
- Fácil de mantener
- Fiabilidad de la solución
- Escalabilidad, en función de futuras necesidades

VII.3. Planificación

Se incluyen a continuación las tareas y, por lo tanto, el alcance a cubrir para las actividades propuestas:

Nombre de tarea	Duración	Comienzo	Fin
Evolucion Proyecto	**109 días**	**mar 01/04/14**	**lun 01/09/14**
Comienzo Proyecto	0 días	mar 01/04/14	mar 01/04/14
Análisis herramientas	**98 días**	**mar 01/04/14**	**jue 14/08/14**
Recopilación información de detectores en el mercado	10 días	mar 01/04/14	lun 14/04/14
Selección detector	9 días	lun 04/08/14	jue 14/08/14
Aplicación	**94 días**	**mar 15/04/14**	**vie 22/08/14**
Desarrollo aplicación	45 días	mar 15/04/14	lun 16/06/14
Pruebas unitarias	5 días	mar 17/06/14	lun 23/06/14
Integración de los detectores en aplicación	20 días	mar 24/06/14	lun 21/07/14
Pruebas con cada detector	9 días	mar 22/07/14	vie 01/08/14
Integración del detector en aplicación	1 día	vie 15/08/14	vie 15/08/14
Pruebas end to end	5 días	lun 18/08/14	vie 22/08/14
Documentación	**11 días**	**vie 15/08/14**	**lun 01/09/14**
Documentación programa	5 días	vie 15/08/14	jue 21/08/14
Documentación pruebas	5 días	lun 25/08/14	vie 29/08/14
Fin proyecto	0 días	lun 01/09/14	lun 01/09/14

Figura 8: Desglose de tareas del proyecto

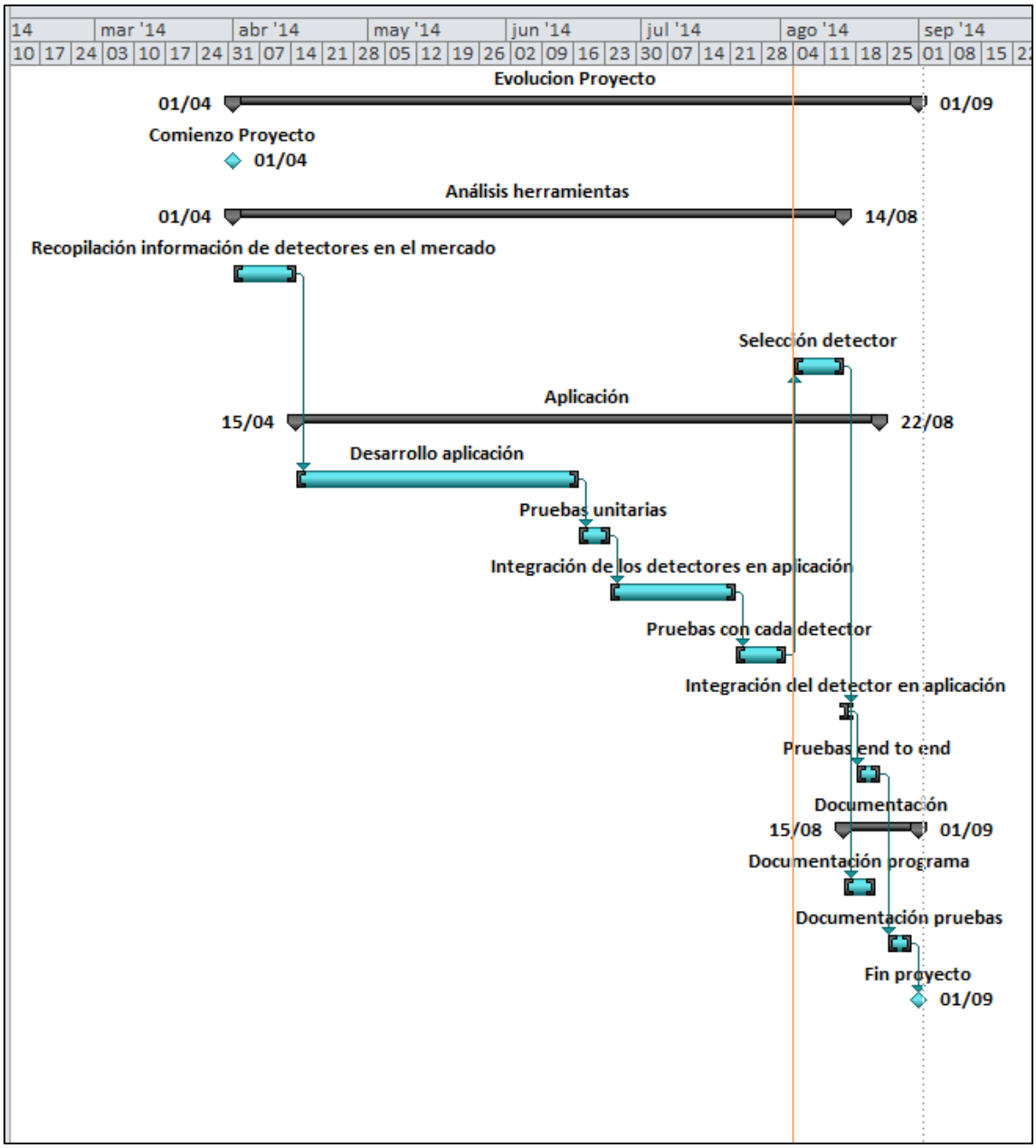


Figura 9: Diagrama Gantt del proyecto

VII.3.1. Recursos

Equipo de trabajo

Todos los recursos asignados al proyecto serán coordinados globalmente por el Proyect Manager. Cada recurso asignado al proyecto tendrá el nivel de conocimientos necesario y la dedicación requerida para conseguir los objetivos del mismo de acuerdo con la calidad establecida y los plazos acordados para el mismo.

El personal ofertado es el que se muestra en la tabla siguiente:

Perfil	Recursos
Project Manager	1
Senior Analyst Programmer	1

Project Manager

Se encarga de dirigir al analista programador para la realización del proyecto asignándole tareas y proveyéndole comentarios de Facebook

Senior Analyst Programmer

Responsable de investigar, diseñar e implementar el software necesario para cumplir los requerimientos del proyecto. Debe ser programador experto en el lenguaje de programación Java para el desarrollo de la aplicación y de Office para poder analizar resultados y presentarlos en el formato correspondiente.

El tiempo total del proyecto ha sido 109 días que equivale a 436 horas.

Materiales Requeridos

En este apartado se desglosan los materiales software y hardware necesarios para la implementación del proyecto:

- Windows 7 Professional 64 bits: Sistema operativo.
- MS Project 2010: Necesario para la gestión del proyecto.
- MS Office 2010:
  - Microsoft Word para la realización de la memoria
  - Microsoft Excel requerido para el análisis de resultados del programa.
- PDFCreator 1.7.3: Se ha utilizado para pasar convertir la memoria a formato pdf.
- Eclipse Kepler: Entorno de desarrollo donde se crea la herramienta e integra con los distintos detectores.

- Internet: Google y redes sociales, abastecimiento de información.
- PC: Portátil HP Pavilion Procesador i7, 8 GB RAM
- Dropbox: Necesario para guardar una copia de seguridad de todos los datos del proyecto, ficheros, memoria, software, código fuente.

VII.4. Valoración económica

VII.4.1 Presupuesto

El coste asociado a los servicios profesionales para la ejecución del proyecto se presenta en la siguiente tabla:

Tipo Recurso	Perfil	Jornadas (horas)	Recursos	Precio Perfil	Precio
Equipo Trabajo	Project Manager	40	1400	35€/h	1400€
	Senior Analyst Programmer	396	9900	25€/h	9900€
				Total	**13.300€**
Materiales	Windows 10 Home				155€
	Office 2013 Home & Business PKC (1 Licencia)				119€
	Project 2010				30€
	PC Portatil				999€
				Total	**1303€**
Total					**14.603€**

VII.4.2 Consideraciones comerciales

El alcance de los servicios ofrecidos está limitado a las tareas y actividades que se describen en este capítulo, quedan fuera del alcance cualquier trabajo no descrito específicamente.

Las condiciones de comerciales se realizarán en arreglo a las siguientes directrices:

- Se emitirá un pedido por el importe total de los trabajos.
- El pago se realizará a los 85 días de la fecha factura.
- Se facturará siguiendo el siguiente plan de facturación:
  - 30% a la aceptación de la Oferta.
  - 70% a la finalización del proyecto.
- El IVA no está incluido en ninguno de los precios ofertados.
- La validez de la oferta es de 30 días.

REFERENCIAS

Capítulo II: Motivación.

- Redes Sociales:
  - http://www.informatica-hoy.com.ar/redes-sociales/La-historia-de-las-redes-sociales.php
  - http://www.informatica-hoy.com.ar/redes-sociales/Que-es-Red-Social.php
  - Ranking top 5 http://www.informacionempleo.com/las-10-redes-sociales-mas-populares/
  - Ventajas http://redespymes.com/2013/12/04/publicidad-en-redes-sociales-ventajas-y-desventajas-para-las-empresas/
  - Evolución http://www.marketingdirecto.com/actualidad/social-media-marketing/breve-historia-de-las-redes-sociales/

- Apps
  - http://www.idgtv.es/entusiastas/habitos-de-uso-de-internet-movil-de-los-jovenes-espanoles
  - http://madrid.theappdate.com/informe-sobre-las-apps-en-espana-2013-asi-usamos-las-apps/#
- Detectores
  - http://publicaciones.uci.cu/index.php/SC/article/viewFile/37/38

Capítulo III: Estado del arte

- Detectores online:
  - Detect Lang
    - http//www.detectlang.com/es/index.html
  - What Language
    - http://whatlanguageisthis.com/
  - Lang Detector
    - http://www.es.langdetector.com/
  - Detector de idiomas
    - http://spark.rstudio.com/gilbellosta/detector_idiomas/
  - Poliglot 3000
    - http://www.polyglot3000.com/es/index.shtml

Capítulo V: Caso Práctico

- Historia King:
  - http://en.wikipedia.org/wiki/King_%28company%29
  - https://king.com/es/
  - http://es.wikipedia.org/wiki/Candy_Crush_Saga
  - http://www.gadwoman.com/2013/09/pet-rescue-y-papa-pear-los-nuevos-juegos-de-los-autores-de-candy-crush/
  - http://www2.prnewswire.com.br/releases/es/kingcom-lanza-juego-en-facebook-pyramid-solitaire-saga/24415
  - http://www.ipadizate.es/2014/07/05/pyramid-solitaire-saga-lanzamiento-94985/
  - http://www.europapress.es/portaltic/videojuegos/noticia-bubble-witch-saga-juego-burbujas-regresa-diseno-menos-malefico-mas-modos-juego-20140521155704.html
  - http://areajugones.es/2014/01/farm-heroes-saga-tambien-disponible-para-android-e-ios/#.VRaSoeFHUUw
  - https://play.google.com/store/apps/details?id=com.king.farmheroessaga&hl=es
  - http://company.king.com/our-games.aspx
  - http://celudescarga.com/descargar-pepper-panic-saga-oficial-para-android/
- Definiciones
  - http://recursostic.educacion.es/observatorio/web/ca/internet/web-20/1043-redes-sociales?start=6

ANEXO I: Textos para análisis capítulo III

Textos en Español

- **Biografía:**

Se inició como periodista en el diario "El Pueblo", donde trabajó doce años. Pasó los primeros años de su carrera periodística a mitad de camino entre los países africanos con conflictos bélicos y las antiguas colonias Españolas (Sahara y Guinea Ecuatorial). Como reportero, Arturo ha cubierto, entre otros conflictos, la guerra de Chipre, diversas fases de la guerra del Líbano, la guerra de Eritrea, la campaña de 1975 en el Sahara, la guerra del Sahara, la guerra de las Malvinas, la guerra de El Salvador, la guerra de Nicaragua, la guerra del Chad, la crisis de Libia, las guerrillas del Sudán, la guerra de Mozambique, la guerra de Angola, el golpe de estado de Túnez, etc. Los últimos conflictos que ha vivido son: la revolución de Rumania (1989-90), la guerra de Mozambique (1990), la crisis y guerra del Golfo (1990-91), la guerra de Croacia (1991) y la guerra de Bosnia (1992-93-94).

Desde 1991 y, de forma continua, escribe una página de opinión en "El Semanal", suplemento del Grupo Correo que se distribuye simultáneamente en 25 diarios españoles, y que se ha convertido en una de las secciones más leídas de la prensa española, superando los 4.500.000 de lectores.

- **Titular Periódico:**

El bulevar de la discordia costará ocho millones de euros.

- **Resumen Periódico:**

Tras ganar al Fuenlabrada, MARCA recordó a Pablo Laso el triste final de sus predecesores en el récord, algo que no altera sus planes: "Los récords no son objetivos que nos hayamos marcado".

- **Canción:**

Este adiós no maquilla un hasta luego,
este nunca no esconde un ojala,
estas cenizas no juegan con fuego,
este ciego no mira para atrás.
Este notario firma lo que escribo,
esta letra no la protestaré,
ahórrate el acuse de recibo,
estas vísperas son las de después.
A este ruido tan huérfano de padre

no voy a permitirle que taladre
un corazón podrido de latir.
Este pez ya no muere por tu boca,
este loco se va con otra loca,
estos ojos no lloran más por ti.

Esta sala de espera sin esperanza,
estas pilas de un timbre que se secó
este helado de fresa de la venganza
esta empresa de mudanza
con los muebles del amor.

Esta campana mora en el campanario,
esta mitad partida por la mitad,
estos besos de Judas, este calvario,
este look de presidiario,
esta cura de humildad.

Este cambio de acera de tus caderas,
estas ganas de nada menos de ti
este arrabal sin grillos en primavera,
ni espaldas con cremalleras,
ni anillos de presumir.

Esta casita de muñecas de alterne
este racimo de pétalos de sal
este huracán sin ojos que lo gobiernen
este jueves, este viernes
y el miércoles que vendrá.

Textos en Inglés

- **Biografía:**

Spanish novelist and journalist , was born on November 25, 1951 in Cartagena. Bachelor of CC. Policies and Journalism. One of the most internationally successful authors in the history of Spanish letters. A boy walking his grandfather 's library looking for adventure stories.

Pérez Reverte was devoted exclusively to literature after living 21 years (1973-1994) as a reporter for newspapers, radio and television, informatively covering international conflicts in that period.

It began as a journalist in the newspaper "The People" , where twelve years worked . He spent the first years of his journalistic career midway between African countries and wars the old Spanish colonies (Sahara and Equatorial Guinea) . As a reporter, Arturo has covered , among other conflicts , the war in Cyprus, various phases of the war in Lebanon , the war in Eritrea , the campaign of 1975 in the Sahara , Sahara War , the Falklands War , war of El Salvador , Nicaragua war , war of Chad, the Libyan crisis , the guerrillas of the Sudan , the war in Mozambique , the war in Angola, the coup in Tunisia , etc. . The last conflict that has lived are: Romania Revolution (1989-90) , the war in Mozambique (1990) , the crisis and Gulf War (1990-91) , the war in Croatia (1991) and the war in Bosnia (1992-93-94) .

Since 1991, continuously, write a review page in "The Week" , Group E supplement that is distributed simultaneously in 25 Spanish newspapers , and has become one of the most read sections of the Spanish press, surpassing 4,500,000 readers..

- **Titular Periódico:**

  Boulevard discord cost eight million euros

- **Resumen Periódico:**

After winning the Fuenlabrada, Pablo Laso MARCA reminded the sad end of his predecessors in the record, which does not alter its plans: "Records are not goals we have set ourselves."

- **Canción:**

This makeup is not bye a see you later ,
this hides a hopefully never ,
these ashes do not play with fire
this blind does not look back.
This notary signature what I write ,
this letter not declare to ,
save your receipt ,
these are the following eve .
In this noise as fatherless
I will not allow you to drill
a rotting heart beating.
This fish never die by your mouth

This goes with another crazy crazy
these eyes do not cry for you.

This waiting without hope,
these piles of dried timbre
This strawberry ice cream Revenge
this mover
Love the furniture .

This bell blackberry on the bell ,
half halved ,
these kisses of Judas , this ordeal ,
this look inmate ,
this humbling .

This change in sidewalk of your hips ,
these feel like anything less than you
this suburb without crickets in spring,
nor backs with zippers
or rings boast.

This dollhouse hostess
this cluster of petals salt
this hurricane without eyes that govern
Thursday , Friday
and Wednesday will

Textos en Portugués

- **Biografía x:**

Romancista e jornalista espanhol , nasceu em 25 de novembro de 1951 em Cartagena. Bacharel em CC . Políticas e Jornalismo . Um dos autores mais internacional bem sucedido na história das letras espanholas. Um menino que anda biblioteca de seu avô à procura de histórias de aventura.

Pérez Reverte foi dedicada exclusivamente à literatura depois de viver 21 anos (1973-1994) como repórter para jornais , rádio e televisão , cobrindo informatively conflitos internacionais nesse período.

Começou como jornalista no jornal " O Povo ", onde trabalhou 12 anos . Ele passou os primeiros anos de sua carreira jornalística no meio do caminho entre os países africanos e as guerras das antigas colônias espanholas (Sahara e Guiné Equatorial) . Como repórter , Arturo cobriu , entre outros conflitos , a guerra no Chipre , várias fases da guerra no Líbano , a guerra na Eritreia , a campanha de 1975 na Sahara , Sahara Guerra , a Guerra das Malvinas , guerra de El Salvador, Nicarágua guerra , Guerra do Chade , a crise da Líbia , os guerrilheiros do Sudão , a guerra em Moçambique , a guerra em Angola , o golpe na Tunísia , etc . O último conflito que viveu são: Revolução Roménia (1989-1990) , a guerra em Moçambique (1990), a crise ea guerra do Golfo (1990-1991) , a guerra na Croácia (1991) e da guerra na Bósnia (1992-93-94) .

Desde 1991, de forma contínua, escrever uma página de revisão de " A Semana " , Grupo E suplemento que é distribuído simultaneamente em 25 jornais espanhóis , e tornou-se uma das seções mais lidas da imprensa espanhola , ultrapassando 4.500.000 leitores.

- **Titular Periódico:**

  Boulevard discórdia custar 8.000.000 €.

- **Resumen Periódico:**

Depois de vencer o Fuenlabrada, Pablo Laso BRAND lembrou o triste fim de seus antecessores no registro, o que não altera seus planos: "Os registros não são metas a que nos propusemos.".

- **Canción:**

Esta composição não é um adeus vê-lo mais tarde ,
este esconde um esperemos que nunca ,
essas cinzas não brincar com fogo
este cego não olha para trás .
Esta assinatura notário o que eu escrevo ,
esta carta não declaram ,
salvar o seu recebimento,
estes são os seguintes véspera .
Neste ruído como órfão
Eu não vou permitir que você para perfurar
um coração batendo em decomposição.
Este peixe nunca morrem por sua boca
Isto vai com outro louco louco

esses olhos não choram por você.

Esta espera sem esperança,
estas pilhas de secas timbre
Este sorvete de morango Revenge
este motor
Ame o mobiliário.

Este blackberry sino em sino ,
metade metade,
esses beijos de Judas , esta provação ,
este detento olhar ,
esta humilhante.

Esta mudança na calçada do seu quadril,
estes se sente como nada menos do que você
este subúrbio sem grilos na primavera,
nem costas com zíperes
ou anéis de se vangloriar.

Este hostess casa de bonecas
este conjunto de pétalas sal
este furacão sem olhos que regem
quinta-feira, sexta-feira
e quarta-feira será

Textos en Alemán

- **Biografía x:**

Spanischer Schriftsteller und Journalist, wurde am 25. November 1951 in Cartagena geboren. Bachelor of CC . Politik und Journalismus . Einer der international erfolgreichsten Autoren in der Geschichte der spanischen Buchstaben. Ein Junge, der Bibliothek seines Großvaters auf der Suche nach Abenteuer-Geschichten .

Pérez Reverte wurde ausschließlich der Literatur nach dem Leben 21 Jahre (1973-1994) als Reporter für Zeitungen , Radio und Fernsehen gewidmet , informativ über internationale Konflikte in diesem Zeitraum.

Es begann als Journalist in der Zeitung " The People " , wo zwölf Jahre lang gearbeitet. Er verbrachte die ersten Jahre seiner journalistischen Laufbahn in der Mitte zwischen afrikanischen Ländern und Kriege die alten spanischen Kolonien (Sahara und Äquatorialguinea) . Als Reporter hat Arturo abgedeckt , unter anderen Konflikten , der Krieg in Zypern, verschiedenen Phasen des Krieges im Libanon, dem Krieg in Eritrea, der Kampagne von 1975 in der Sahara , Sahara Krieg , Falkland-Krieg , Krieg von El Salvador, Nicaragua Krieg , Krieg des Tschad, Libyen-Krise , die Guerilla des Sudan, den Krieg in Mosambik, den Krieg in Angola, den Putsch in Tunesien , etc.. Der letzte Konflikt, der gelebt hat, sind : Rumänien Revolution (1989-1990) , der Krieg in Mosambik (1990), die Krise und Golfkrieg (1990-1991), der Krieg in Kroatien (1991) und der Krieg in Bosnien (1992-93-94) .

Seit 1991 kontinuierlich , schreiben Sie eine Kritik Seite in "The Week" , Gruppe E Ergänzung, die gleichzeitig in 25 spanischen Zeitungen verteilt wird , und hat sich zu einem der meist gelesenen Abschnitte der spanischen Presse und übertraf damit 4,5 Millionen Leser.

- **Titular Periódico:**

  Boulevard Zwietracht kostet 8.000.000 €

- **Resumen Periódico:**

Nach dem Gewinn des Fuenlabrada, erinnerte Pablo Laso BRAND das traurige Ende seiner Vorgänger in der Aufzeichnung, die ihre Pläne nicht ändert: "Rekorde sind nicht Ziele, die wir uns selbst gesetzt haben."

- **Canción:**

Das Make-up ist nicht ein Wiedersehen bis später ,
Diese verbirgt sich eine hoffentlich nie ,
diese Asche nicht mit dem Feuer spielen
dieser blinde schaut nicht zurück .
Dieser Notar Unterschrift , was ich schreibe ,
dieser Brief nicht zu erklären ,
speichern Sie Ihre Quittung ,
Dies sind die folgenden Vorabend .
In diesem Geräusch, als Waisen
Ich werde nicht zulassen, dass Sie zu bohren
ein faul Herz schlagen .
Dieser Fisch nie durch den Mund sterben
Dies geht mit einem anderen crazy crazy

diese Augen nicht für Sie weinen.

Diese Warte ohne Hoffnung,
diese Haufen von getrockneten
Das Erdbeereis Revenge
Diese Maschine
Die Liebe der Möbel.

Diese Glocke Blackberry auf der Glocke ,
halb halbiert ,
diese Küsse des Judas, dieser Tortur ,
dieser Blick Häftling ,
Diese demütigende .

Diese Änderung in der Bürgersteig der Hüfte,
diese fühlen sich wie nichts weniger, als Sie
Dieser Vorort ohne Grillen im Frühjahr,
Rücken noch mit Reißverschlüssen
oder Ringe rühmen.

Dieses Puppenhaus Gastgeberin
dieser Cluster der Blütenblätter Salz
dieser Hurrikan ohne Augen , die regieren
Donnerstag, Freitag
und Mittwoch wird

Textos en Francés

- **Biografía x:**

Romancier et journaliste espagnol , est né le 25 Novembre 1951 à Cartagena . Baccalauréat en CC . Politiques et journalisme . L'un des auteurs les plus internationalement réussi dans l'histoire des lettres espagnoles . Un garçon marche dans la bibliothèque de son grand-père à la recherche de récits d'aventures .

Pérez Reverte a été consacrée exclusivement à la littérature après avoir vécu 21 années (1973-1994) en tant que journaliste pour les journaux , la radio et la télévision , couvrant informative conflits internationaux dans cette période .

Il a commencé en tant que journaliste dans le journal " The People" , où douze années travaillé . Il a passé les premières années de sa carrière journalistique à mi-chemin entre les pays africains et les guerres anciennes colonies espagnoles (Sahara et Guinée équatoriale) . En tant que journaliste , Arturo a couvert , entre autres conflits , la guerre à Chypre , les différentes phases de la guerre au Liban , la guerre en Érythrée , la campagne de 1975 dans le Sahara , la guerre du Sahara , la guerre des Malouines , la guerre d'El Salvador , Nicaragua guerre , la guerre du Tchad , la crise libyenne , les guérilleros du Soudan , la guerre au Mozambique , la guerre en Angola , le coup d'Etat en Tunisie , etc . Le dernier conflit qui a vécu sont : Roumanie Révolution (1989-1990) , la guerre au Mozambique (1990) , la crise et la guerre du Golfe (1990-1991) , la guerre en Croatie (1991) et la guerre en Bosnie (1992-93-94) .

Depuis 1991 , en continu , écrire une page de révision dans " La Semaine " , Groupe E supplément qui est distribué simultanément dans 25 journaux espagnols , et est devenu l'une des sections les plus lus de la presse espagnole , dépassant 4,5 millions de lecteurs .

- **Titular Periódico:**

  Boulevard discorde coûte € 8.000.000.

- **Resumen Periódico:**

Après avoir remporté le Fuenlabrada, Pablo Laso MARQUE rappelé la triste fin de ses prédécesseurs dans le dossier, qui ne modifie pas ses plans: «Les records ne sont pas des objectifs que nous nous sommes fixés."

- **Canción:**

Ce maquillage n'est pas un bye -vous voir plus tard,
ce cache un espérons jamais ,
ces cendres ne jouent pas avec le feu
ce store ne regarde pas en arrière .
Cette signature notaire ce que j'écris ,
cette lettre ne déclarent ,
enregistrer votre réception ,
ceux-ci sont à la veille suivante.
En ce bruit comme orphelin
Je ne vais pas vous permettre de percer
un battement de coeur pourriture .
Ce poisson ne meurent jamais par la bouche
Cela va avec un autre fou fou

ces yeux ne pleurent pas pour vous.

Cette attente sans espoir ,
ces piles de timbre séché
Cette crème glacée Revenge fraise
ce moteur
Aimez les meubles .

Cette mûre de cloche sur la cloche ,
moitié moitié,
ces baisers de Judas , cette épreuve ,
ce détenu de regard ,
cette leçon d'humilité.

Ce changement dans le trottoir de vos hanches ,
ceux-ci se sentent comme rien de moins que vous
cette banlieue sans grillons au printemps ,
ni le dos avec fermeture éclair
ou anneaux offrent .

Cette hôtesse de maison de poupée
ce groupe de pétales de sel
cet ouragan sans yeux qui régissent
jeudi, vendredi
et mercredi

Textos en Italiano

- **Biografía x:**

Romanziere spagnolo e giornalista , è nato il 25 novembre 1951 a Cartagena . Bachelor di CC . Politiche e Giornalismo . Uno dei più internazionale riuscita autori nella storia delle lettere spagnole . Un ragazzo che cammina biblioteca di suo nonno in cerca di storie d'avventura .

Pérez Reverte è stato dedicato esclusivamente alla letteratura dopo aver vissuto 21 anni (1973-1994) come reporter per giornali, radio e televisione , informatively copre conflitti internazionali in quel periodo .

E 'iniziato come giornalista nel quotidiano "Il Popolo ", dove dodici anni lavoravano . Ha trascorso i primi anni della sua carriera giornalistica a metà strada tra i paesi africani e le guerre le vecchie colonie spagnole (Sahara e Guinea Equatoriale) . Come giornalista , Arturo ha ricoperto , tra gli altri conflitti , la guerra di Cipro , varie fasi della guerra in Libano , la guerra in Eritrea , la campagna del 1975 nel Sahara , Sahara guerra , la guerra delle Falkland , guerra di El Salvador , Nicaragua guerra , guerra del Ciad , la crisi libica , i guerriglieri del Sudan , la guerra in Mozambico , la guerra in Angola , il colpo di stato in Tunisia , ecc . L' ultimo conflitto che ha vissuto sono : Romania Revolution (1989-1990) , la guerra in Mozambico (1990) , la crisi e la guerra del Golfo (1990-1991) , la guerra in Croazia (1991) e la guerra in Bosnia (1992-93-94) .

Dal 1991 , ininterrottamente , scrivere una pagina di commento in "The Week" , Gruppo E supplemento che viene distribuito simultaneamente in 25 giornali spagnoli , ed è diventato una delle sezioni più letti della stampa spagnola , superando 4.500.000 lettori.

- **Titular Periódico:**

  Boulevard discordia è costato € 8.000.000

- **Resumen Periódico:**

Dopo aver vinto il Fuenlabrada, Pablo Laso BRAND ricorda la triste fine dei suoi predecessori nel record, che non modifica i suoi piani: "I record non sono obiettivi che ci siamo prefissati."

- **Canción:**

Questo trucco non è un addio ci vediamo più tardi ,
questo nasconde un si spera mai ,
queste ceneri non giocare con il fuoco
questo cieco non guardare indietro .
Questa firma notaio quello che scrivo ,
questa lettera non dichiarano ,
salvare la ricevuta ,
questi sono i seguenti vigilia .
In questo rumore come orfani
Non ti permetterò di drill
un cuore che batte in decomposizione .
Questo pesce non muoiono mai dalla tua bocca
Questo va con un altro pazzo pazzo

questi occhi non piangono per te .

Questo attesa senza speranza,
queste pile di secchi timbro
Questo gelato alla fragola Revenge
questo mover
Adoro i mobili .

Questo mora campana sulla campana ,
metà dimezzato ,
questi baci di Giuda, questo calvario ,
questo sguardo detenuto ,
questo umiliante .

Questo cambiamento di marciapiede dei fianchi ,
questi si sentono come qualcosa di meno di quello che
questo sobborgo , senza grilli in primavera ,
né spalle con zip
o anelli vantano .

Questo hostess dollhouse
questo gruppo di petali di sale
questo uragano senza occhi che governano
Giovedi, Venerdì
e Mercoledì sarà

ANEXO II: Código del programa

En este Anexo II se detalla el código de la herramienta de detección de idiomas seleccionada.

Como se explicó en el capítulo IV la estructura del proyecto consta de 4 métodos:

- Main: es el método principal desde el cual se ejecuta el programa, tiene todas las variables globales y necesarias para la correcta ejecución.
- Leer: se encarga de leer la estructura de ficheros de los post en la ruta indicada, así como obtener dentro de cada post los mensajes que han escrito los usuarios en la red social
- Escribir: método que guarda en fichero (en la ruta de salida) la siguiente información, nombre de carpeta, nombre del post, mensaje del usuario y el idioma del mensaje. Con el fichero generado de salida permite posteriormente analizar los resultados.
- Identificar Idioma: este método es el encargado de integrar el detector del lenguaje con el programa.

1. Método Main

```
import java.io.File;
import java.io.FileWriter;
import java.io.PrintWriter;
import java.util.HashMap;
import java.util.Set;


public class ComentariosRedesSociales2 {

	public static void main(String[] args) {
		// Variables
		String ruta = "C:/Users/Estela/workspace/textos";
		String rutaResultados = "C:/Users/Estela/workspace/textos/Resultados2.txt";

		HashMap<String, File> listadoFicherosComentarios;

		String textosCompletos = null;
		String mensaje = null;
		String[] mensajeTextos = null;

		FileWriter ficheroResultados = null;
		PrintWriter pw = null;
```

```
// Ejecucion programa
try {

	// Objeto para llamar al detector de lenguajes
	IdentificarIdioma2 idiomaComentarios = new IdentificarIdioma2();

	// Cargo los ficheros
	LeerFicheros2 leer = new LeerFicheros2();
	File[] ficheros = new File[1];
	ficheros[0] = new File(ruta);

	// Ficheros para guardar resultados
	EscribirFicheros2 resultados = new EscribirFicheros2();
	ficheroResultados = new FileWriter(rutaResultados);
	pw = new PrintWriter(ficheroResultados);
	pw.println("nombreCarpeta" + "Ç" + "nombrePost" + "Ç" + "mensaje" + "Ç" +
"idioma");

	// Recorro todos los ficheros, detecto el idioma y lo guardo en un fichero
	listadoFicherosComentarios =
leer.generarListadoFicherosComentarios(ficheros);
	Set<String> keys = listadoFicherosComentarios.keySet();

	for (String key : keys) {
		File auxFile = listadoFicherosComentarios.get(key);
		textosCompletos = leer.ficherosTextos(auxFile);

		if (textosCompletos == null) {

			//Compruebo si el fichero está vacio y si es así no continuo
voy al siguiente fichero

			continue;
		}

		//Para cada post extraigo todos los mensajes y los analizo
		mensajeTextos = leer.extraerTodosMesajes(textosCompletos);

		for (int i = 0; i < mensajeTextos.length; i++) {

			mensaje = leer.extraerMesaje(mensajeTextos[i]);

			if (mensaje != null) {

				String idioma =
idiomaComentarios.detectaIdioma(mensaje);
				resultados.nuevoFichero(pw,
auxFile.getParentFile().getParentFile().getName(), auxFile.getParentFile().getName(), mensaje,
idioma);

			}

		}

	}

} catch (Exception e2) {
```

```
                    // Catch de excepciones Ficheros
                    e2.printStackTrace();

                }

                finally {
                    try {
                        // Cerramos el fichero una vez finalizadas todas las iteraciones.
                        if (null != ficheroResultados)
                            ficheroResultados.close();
                    } catch (Exception e2) {
                        e2.printStackTrace();
                    }

                }


        }

}
```

2 Método EscribirFicheros

```
import java.io.PrintWriter;


public class EscribirFicheros2 {

        //Metodo para escribir en fichero los resultados
        public void nuevoFichero(PrintWriter pw, String nombreCarpeta,String nombrePost, String
mensaje, String idioma) {

                try {

                    pw.println(nombreCarpeta + "Ç" + nombrePost + "Ç" + mensaje + "Ç" +
idioma);

                } catch (Exception e) {
                    e.printStackTrace();
                }
        }


}
```

3 Método IdentificarIdioma

```
import me.champeau.ld.UberLanguageDetector;


public class IdentificarIdioma2 {

	//Metodo para detectar el idioma del texto
	public String detectaIdioma(String text)
	{
		UberLanguageDetector detector = UberLanguageDetector.getInstance();
		String language = detector.detectLang(text);

		return language;

	}


}
```

4. Método LeerFicheros

```
import java.io.BufferedReader;
import java.io.DataInputStream;
import java.io.File;
import java.io.FileInputStream;
import java.io.IOException;
import java.io.InputStreamReader;
import java.util.HashMap;


public class LeerFicheros2 {

	//Metodo que recorre la estructura de carpetas y de cada directorio se queda con el nombre del post y el Fichero
	public HashMap<String, File> generarListadoFicherosComentarios(File[] ficheros) {

		HashMap<String, File> mapTextos = new HashMap<String, File>();

		for (File file : ficheros) {

			if (file.isDirectory()) {

				// Llamamos la función de manera recurrente para recorrer todos los ficheros
```

```
                                HashMap<String, File> mapAux =
generarListadoFicherosComentarios(file.listFiles());

                                // Añado cada uno de los ficheros encontrados
                                mapTextos.putAll(mapAux);

                        } else {

                                File auxFile = new File(file.getParent());
                                mapTextos.put(auxFile.getName(), file);

                        }
                }

                return mapTextos;

        }

        //Metodo que dado un fichero devuelve en un string el texto que contiene
        public String ficherosTextos(File fichero) {
                DataInputStream entrada = null;
                String textoLeido = "";

                try {

                        // Creamos el objeto de entrada
                        FileInputStream fis = new FileInputStream(fichero);
                        entrada = new DataInputStream(fis);

                        // Creamos el Buffer de Lectura
                        BufferedReader buffer = new BufferedReader(new
InputStreamReader(entrada));

                        // Leer el archivo linea por linea
                        String linea;
                        do{
                                linea = buffer.readLine();
                                textoLeido += linea;

                        } while(linea != null);

                } catch (Exception ex) {

                        // Catch de excepciones
                        System.err.println("Ocurrio un error: " + ex.getMessage());

                } finally {

                        // Cerramos el archivo
                        try {

                                entrada.close();

                        } catch (IOException e1) {

                                e1.printStackTrace();
```

```
                }

            }

            return textoLeido;

        }

        //Metodo que dado un post con todos los comentarios los separa y guarda cada comentario en
una posicion del array
        public String [] extraerTodosMesajes(String comentariosCompletos) {

            String [] arraySplit = comentariosCompletos.split("like_count");

            for (String s : arraySplit){

                this.extraerMesaje(s);

            }


            return arraySplit;

        }

        //Metodo que dado un comentario extrae el mensaje
        public String extraerMesaje(String comentarios) {

            String linea = comentarios;
            String mensaje=null;
            String[] arraySplit = linea.split("message:\t");

                if (arraySplit.length > 1) {

                    mensaje = arraySplit[1];

                    if (mensaje != null){

                        mensaje = mensaje.split("created_time")[0];

                    }

                }

            return mensaje;

        }


}
```